Plant Physiology Preview. Published on January 11, 2016, as DOI:10.1104/pp.15.01281

Running title:

Label-based comparative co-expression

Title:

FamNet: A framework to identify multiplied modules driving pathway expansion in plants.

Colin Ruprecht[1], Amelie Mendrinna[1,2], Takayuki Tohge[1], Arun Sampathkumar[3], Sebastian Klie[1], Alisdair R. Fernie[1], Zoran Nikoloski[1], Staffan Persson[1,2,4#], Marek Mutwil[1*#]

[1] Max-Planck-Institute for Molecular Plant Physiology, Am Muehlenberg 1, 14476 Potsdam, Germany

[2] School of Biosciences, the University of Melbourne, Parkville 3010, Victoria, Australia

[3] Division of Biology and Biological Engineering 156-29, California Institute of Technology, Pasadena, CA, 91125, USA

[4] ARC Centre of Excellence in Plant Cell Walls, School of Biosciences, the University of Melbourne, Parkville 3010, Victoria, Australia

#Co-senior authorship

*Corresponding author:

Marek Mutwil

Max-Planck-Institute for Molecular Plant Physiology,

Am Muehlenberg 1,

14476 Potsdam,

Germany

Email: Mutwil@mpimp-golm.mpg.de

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (http://www.plantphysiol.org/site/misc/ifora.xhtml) is: Marek Mutwil (mutwil@mpimp-golm.mpg.de).

This work was supported by the Max-Planck Gesellschaft (CR, AM, TT, ARF, SK, ARF, ZN, MM), R@MAP grant as part of his Professorship at University of Melbourne (SP), the European Commission's Directorate General for Research within the 7th Framework Program (FP7/2007-2013) under grant agreement 270089 (MULTIBIOPRO; CR, TT, ARF and SP) and ERA-CAPS grant EVOREPRO (MM).

M.M. and S.P. conceived the project, C.R. and A.M. performed the knock-out experiments, T.T. and A.R. performed the metabolomic analysis, S.K., Z.N. and M.M. conceived the bioinformatical analyses, M.M. performed the bioinformatical analyses, M.M., S.P., A.S. and C.R. wrote the article with help from all authors.

Abstract

Gene duplications generate new genes that can acquire similar but often diversified functions. Recent studies of gene co-expression networks have indicated that not only genes, but also pathways can be multiplied and diversified to perform related functions in different parts of an organism. Identification of such diversified pathways, or modules, is needed to expand our knowledge of biological processes in plants and to understand how biological functions evolve. However, systematic explorations of modules remain scarce and no user-friendly platform to identify them exists. We have established a statistical framework to identify modules and show that approximately one third of the genes of a plant's genome participate in hundreds of multiplied modules. Using this framework as a basis, we implemented a platform that can explore and visualize multiplied modules in co-expression networks of eight plant species. To validate the usefulness of the platform, we identified and functionally characterized pollen and root specific cell wall modules that multiplied to confer tip-growth in pollen tubes and root hairs, respectively. We, furthermore, identified multiplied modules involved in secondary metabolite synthesis and corroborated them by metabolite profiling of tobacco tissues. The interactive platform, referred to as FamNet is available at http://www.gene2function.de/famnet.html.

Introduction

Transcriptionally associated genes tend to be involved in related biological processes (Usadel et al., 2009). Transcriptional associations, termed co-expression, have been used extensively to infer gene functions in many model organisms (Itkin et al., 2013; Persson et al., 2005; Stuart et al., 2003; Yu et al., 2003). Several web-based tools have

been developed to allow users to exploit such relationships (e.g. (Lee et al., 2015; Mutwil et al., 2010; Obayashi et al., 2011)). Some of these tools offer the possibility to extend the analyses to species that only recently have emerged as tractable systems for genetic engineering, e.g. several plant crop species (Ficklin and Feltus, 2011; Movahedi et al., 2011; Mutwil et al., 2011; Tzfadia et al., 2012). Co-expression patterns may also be conserved across species barriers (Bergmann et al., 2004; Stuart et al., 2003). Such conserved co-expressed patterns can be used to transfer knowledge obtained from a well investigated model species to other organisms, e.g. crop plants, as is possible via several web-tools (Mutwil et al., 2011; Park et al., 2013; Ruprecht et al., 2011; Tzfadia et al., 2012). Furthermore, conserved co-expression patterns tend to be enriched for biologically relevant relationships and can be used to improve predictions (Hansen et al., 2014; Movahedi et al., 2011),

Generally, scientists apply classification schemes to associate gene products with functions. For example, genes and proteins may be associated with a family, a metabolic pathway, subcellular localization, and a protein complex. These classification schemes make it possible to define biological hierarchies and to communicate advances within specific research fields. While classifications are instructive for gene products that are associated to known biological functions, they do not allow for inferences of genes and proteins that lack functional description. Co-expressed gene neighbourhoods, as functional biological units, can associate uncharacterized genes to biological functions (Aoki et al., 2007; Heyndrickx and Vandepoele, 2012; Kanehisa et al., 2015; Langfelder and Horvath, 2008). Prominent examples where this approach has been used include: primary and secondary wall cellulose production (Brown et al., 2005;

93 Persson et al., 2005; Ruprecht et al., 2011), and secondary metabolite production (Itkin
94 et al., 2013; Tohge et al., 2007; Yonekura-Sakakibara et al., 2008) in plants, as well as
95 cholesterol biosynthesis (Langfelder and Horvath, 2008) and cell proliferation (Shi et al.,
96 2010) in mouse and human breast carcinoma, respectively.

97 Recently, several reports have touched upon the notion that related co-
98 expressed gene neighborhoods appear multiple times in an organism. For instance, the
99 primary wall cellulose synthesis neighborhood contains several genes for which close
100 homologs appear in the secondary wall cellulose synthesis neighborhood (Ruprecht et
101 al., 2011). Similarly, a co-expressed gene neighborhood in Arabidopsis is responsible
102 for a specialized phenolic pathway during pollen development (Matsuno et al., 2009),
103 and genes in this neighborhood have close homologs that form co-expressed
104 neighborhoods that partake in phenolic pathways in other parts of the plant (Ehlting et
105 al., 2008). This suggests that co-expressed gene neighborhoods have been duplicated,
106 or even multiplied, and sub- or neo-functionalized during evolution. We refer to such
107 multiplied gene neighborhoods as multiplied modules. A major obstacle to identify
108 multiplied modules has been to label the genes in an appropriate manner. Multiplication
109 of modules was investigated in yeast (Conant and Wolfe, 2006; He and Zhang, 2005;
110 Wapinski et al., 2007), where genes across the whole genome were grouped into
111 families as an indicator of functional relatedness. However, genes from different families
112 might harbour the same protein domains that have analogous functions (Kummerfeld
113 and Teichmann, 2005), and consequently, using only gene families might not detect
114 functionally related modules. Proteins can be labelled by protein domains via the Pfam
115 database (Punta et al., 2012) and through families, for example via the PLAZA

116 database (Proost et al., 2009). An alternative route may therefore be to use both
117 families and domains to label gene products, with the aim to detect multiplied modules.
118 To capture plant-specific modules that might have related functions, our method
119 combines protein domain and gene family labels. We used these labels and developed
120 a statistical pipeline, which detected hundreds of multiplied modules. Furthermore, we
121 established a web-tool, FamNet, that allows the user to retrieve conserved and
122 multiplied modules across and within eight plant species. We used FamNet to identify,
123 and functionally characterize, multiplied modules involved in secondary metabolism and
124 in cell wall biosynthesis in tip growing cells. Our findings suggest that multiplied
125 modules indeed may perform related, but specialized, functions.
126
127 **Results and Discussion**
128 **A statistical pipeline to detect multiplied modules**
129 We have shown that several homologous gene pairs are present in the co-expressed
130 gene neighborhoods of primary and secondary wall cellulose synthesis, respectively
131 (Ruprecht et al., 2011). This discovery led to the question; “Are homologous genes
132 typically found in multiple co-expressed neighborhoods, or modules, and if so, how can
133 we detect such modules?”. Attempts to identify gene modules based on co-expression
134 networks have often been based on clustering algorithms that produce different clusters
135 depending on the network properties and parameter settings (Mao et al., 2009). Here,
136 we developed a statistical pipeline to systematically detect co-expressed gene
137 neighborhoods with common PLAZA gene families and Pfam protein domains labels
138 within and across the eight plant species: *Arabidopsis thaliana*, *Oryza sativa*, *Medicago*
139 *truncatula*, *Populus tremula*, *Hordeum vulgare*, *Glycine max*, *Nicotiana tabacum* and

Triticum spp. (see Supplementary Material for details). Our pipeline consists of two main parts: (i) identification of conserved transcriptional associations of gene family and protein domain labels and (ii) mapping of these conserved associations onto co-expressed gene neighborhoods to find multiplied neighborhoods in genome-wide co-expression networks. These similar gene neighborhoods were then termed gene modules.

The assumptions behind the first part of the pipeline are that functionally related labels, i.e. gene families and Pfam domains, should be co-expressed, and that the co-expression relationships should be conserved across species. We assigned the labels to genes, and any gene can therefore be associated with multiple labels. While the labels used in this study are sequence-based, our pipeline allows inclusion of any type of labels, such as ontology, protein structure information and others. To identify transcriptional association of labels, we transformed co-expressed gene neighborhoods into label co-expression networks (Fig. 1A and B; Supplementary material section 1.1). We then permuted the gene-label assignments to obtain associated labels in the eight plants (Fig. 1C, Supplementary material section 1.2). As conserved co-expression relationships are better estimates for true biological relationships (Hansen et al., 2014; Heyndrickx and Vandepoele, 2012; Mutwil et al., 2011), we only retained co-expressed label associations found in at least two species to assure robustness of the associations (Fig. 1D, Supplementary material section 1.3). We termed the conserved label association network Ensemble Label Association network, or ELA network (available as Supplementary Data 2). The ELA represents conserved associations between gene families and protein domains and can reveal functional associations between these

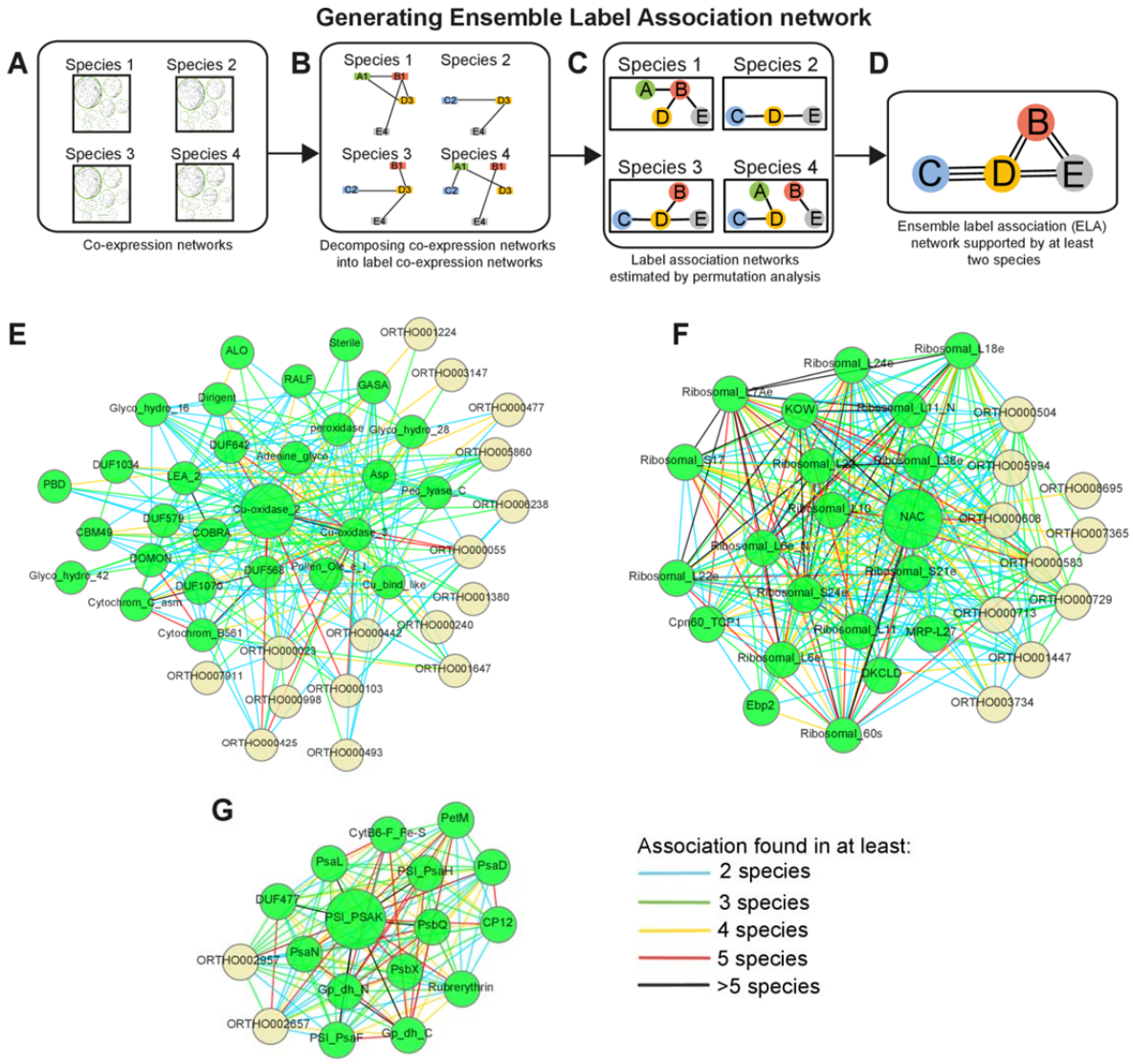

Figure 1. Generating Ensemble Label Association (ELA) network. (A) Co-expression networks derived from the PlaNet platform are used as input. Each gene may be assigned multiple labels. (B) The gene co-expression networks are decomposed into label co-expression networks, where nodes represent labels assigned to genes, and edges represent co-expression relationships between the labels. (C) Associations between labels are detected for each species by label-based node cover and permutation test. The result is label association networks, where nodes represent labels and edges represent associations between the labels. (D) Label association networks are combined into ensemble label association (ELA) network. The number of edges (associations) that are conserved across the different species is determined. In this example, labels C and D are connected in three species (species 2, 3 and 4). (E) ELA of Cu-oxidase_2. (F) ELA of NAC. (G) ELA of PSI_PSAK. Green and yellow nodes represent Pfam and PLAZA labels, respectively. Edges show in how many species an association was found.

labels. Fig. 1E -G shows three ELA regions specific to labels involved in cell wall biosynthesis, photosynthesis and ribosome biogenesis. The ELA region of Cu-oxidase_2 label associated with lignin production during cell wall formation identified

several other labels involved in cell wall biosynthesis, such as COBRA, DUF579 and various carbohydrate-active enzymes (CBMs, glycosyl hydrolases and others, Fig. 1E, (Ruprecht et al., 2011)). The ELA region of nascent polypeptide associated complex (NAC) contains labels that are structural components of ribosomes, ribosome assembly and translation factors (Ebp2, MRP-L27, Cpn60_TCP1, Fig. 1F). Another example of photosystem 1 label PSI_PSAK revealed other components of the photosystem, such as photosystem I (PSI, PSA labels) and photosystem II (PSB labels, Fig. 1G). This part of the pipeline therefore established conserved label associations across eight plant species. The ELA is used to define valid labels when estimating similarities of modules by only using label combinations found in ELA, as described below.

Next, we mapped the conserved label associations (ELA) to the gene co-expression network with the aim to find modules. Importantly, we removed genes that were not supported by the ELA network as they represented non-conserved associations (Fig. 2A and B, Supplemental material section 2.1). As genes in our pipeline can be associated with multiple labels, it is likely that neighborhood similarities are overestimated if only the number of shared labels is used for counting. For example, simple label counting would return the same result when comparing two neighborhoods if (i) each contain one gene with labels ABC, or (ii) each contain three genes with single labels D, E and F. While both examples indicate three labels in common for the neighborhoods, the outcome of (i) is due to the number of labels assigned to the genes. To avoid this potential bias we iteratively binned genes that were associated with the same labels into what we refer to as *label co-occurrences* (Fig. 2C, Supplemental material section 2.2). Label co-occurrences were subsequently counted and used to

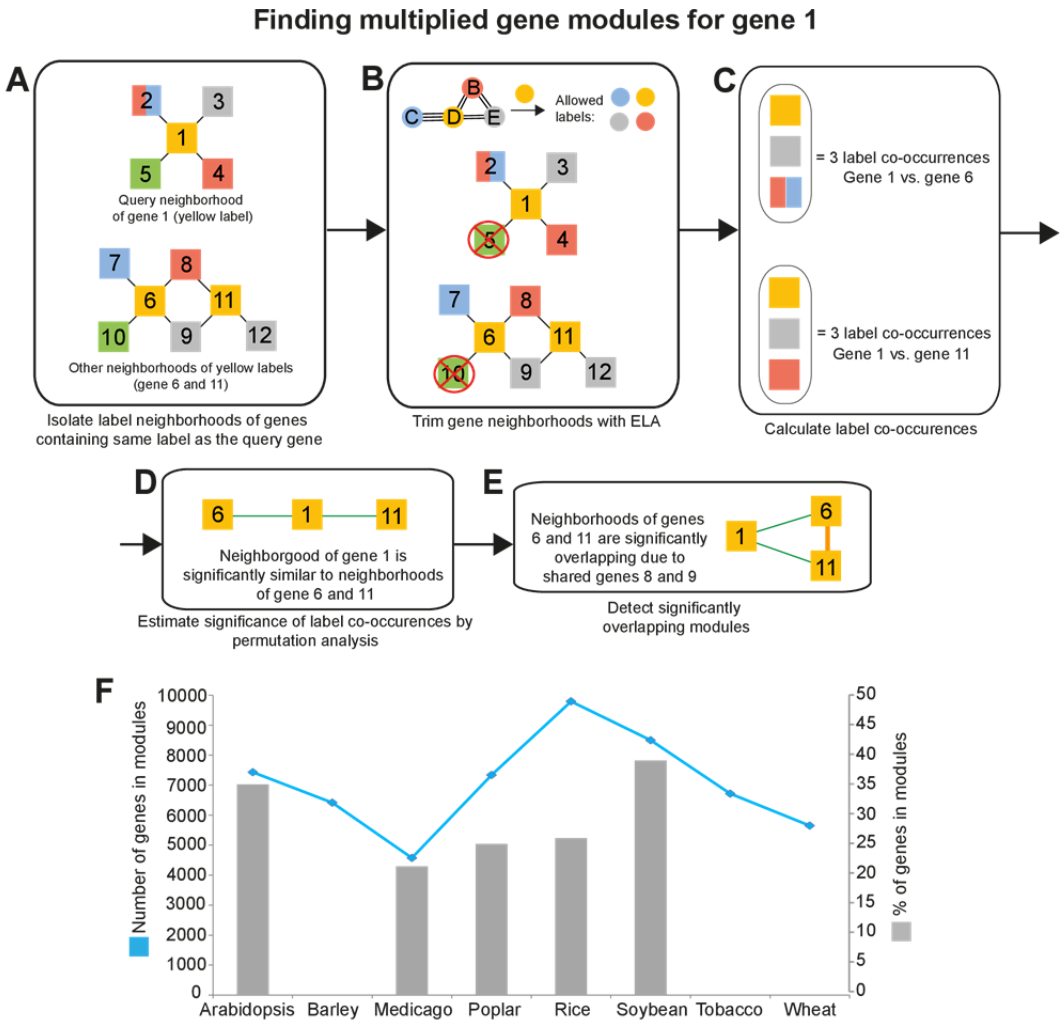

Figure 2. Detecting similar modules. The pipeline is exemplified by searching for similar modules to the neighborhood of gene 1. (A) The neighborhood of the query gene 1 is first isolated. Nodes represent genes, edges represent co-expression relationships and node colors indicate labels found in collected genes. Note that gene 2 has two labels, red and blue. Label neighborhoods of genes containing orange label (genes 6 and 11) are isolated. (B) The neighborhoods are trimmed with ELA, where labels not supported by ELA are removed (Figure 1D). (C) Label co-occurrences found between the neighborhood of the query gene and label neighborhoods are calculated. As gene 2 contains two labels, genes 7 and 8 are collapsed into one label co-occurrence. (D) The significance of found label co-occurrences is estimated by permutation analysis. Green edges indicate similar neighborhoods. (E) Overlapping modules are identified. (F) Total number and percentage of genes assigned to similar modules. Blue line (left y-axis) denotes number of genes assigned to modules. Gray bars (right y-axis) represent the percentage of total genes found on the microarrays that are assigned to modules (right y-axis). Note that % of genes for barley, wheat and tobacco is missing due to the lack of comprehensive genome annotation of the microarrays.

represent neighbourhood similarities (Supplemental material section 2.2). To test which neighborhood pairs are significantly similar, we permuted gene-label associations 1000 times to estimate empirical p-value for each pair (Fig. 2D, See supplemental material

192 section 2.2). Gene neighbourhood pairs that were significantly similar (P<0.01) were
193 then referred to as multiplied modules (Fig. 2E, See supplemental material section 2.3).
194 We note that since label co-occurrences greedily bin genes that have at least one
195 protein domain or gene family in common into one unit, the metric tends to
196 underestimate the similarity of modules. The multiplied modules are available as
197 Supplementary Data 3. Since many of these multiplied modules are overlapping in the
198 co-expression network, we selected only non-overlapping modules in a last step of the
199 pipeline (Fig. 2F, Supplementary Material Section 2.4).
200
201 **Genome wide analysis of multiplied gene modules**
202 We found that between 4,000 (Medicago, blue line in Fig. 2G) and 10,000 (rice, blue
203 line in Fig. 2G) genes were associated with multiplied modules in the eight plants. This
204 indicates that between 22% (Medicago, grey bars in Fig. 2G) to 38% (soybean, grey
205 bars in Fig. 2G) of the genes in the genome of these species were part of the multiplied
206 modules. These numbers are likely to be underestimates as not all genes are
207 represented on microarrays; typically around 60% of the total genes in the genome of
208 these species have corresponding probesets (Mutwil et al., 2011). Also, not all cell
209 types and tissues are covered by the expression data. For example, Medicago lacks
210 microarrays capturing transcriptome of pollen (Mutwil et al., 2011), and consequently,
211 pollen specific modules will not be detected in our study. Finally, since we only
212 considered conserved label associations, the analysis disregards multiplied gene
213 modules that are species specific. Nevertheless, our analysis revealed that a substantial
214 portion of the genes in the eight plant genomes partake in the multiplied modules.

Next, we investigated the module sizes, i.e. how many label co-occurrences any two multiplied modules have in common (Fig. 3). As our pipeline does not use clustering, but is based on neighborhoods, it is possible that some genes of one module also are included in another module. To estimate the number of non-overlapping modules we used a greedy heuristic based on sorting pairs of duplicated modules according to the number of label co-occurrences in descending order, and collected the values of label co-occurrences when modules do not overlap with already collected modules (Supplementary Fig. 1). While this heuristic favors selection of large modules, we found that ~80% of the multiplied modules were small, i.e. similar due to two to five common labels. However, we also identified modules that contained over 15 label co-occurrences (Fig. 3A, Supplementary Data 4), exemplified by large multiplied modules involved in defense response in soybean (Fig. 3B), chromatin remodeling in rice (Supplementary Fig. 2), and ribosome biogenesis in tobacco (Supplementary Fig. 3). This demonstrates that large functionally related modules have been multiplied.

We also investigated the number of times the modules can be multiplied, termed *module degree*. Since some modules are overlapping, we again used a greedy heuristic to select non-overlapping modules by sorting each module according to the degree in

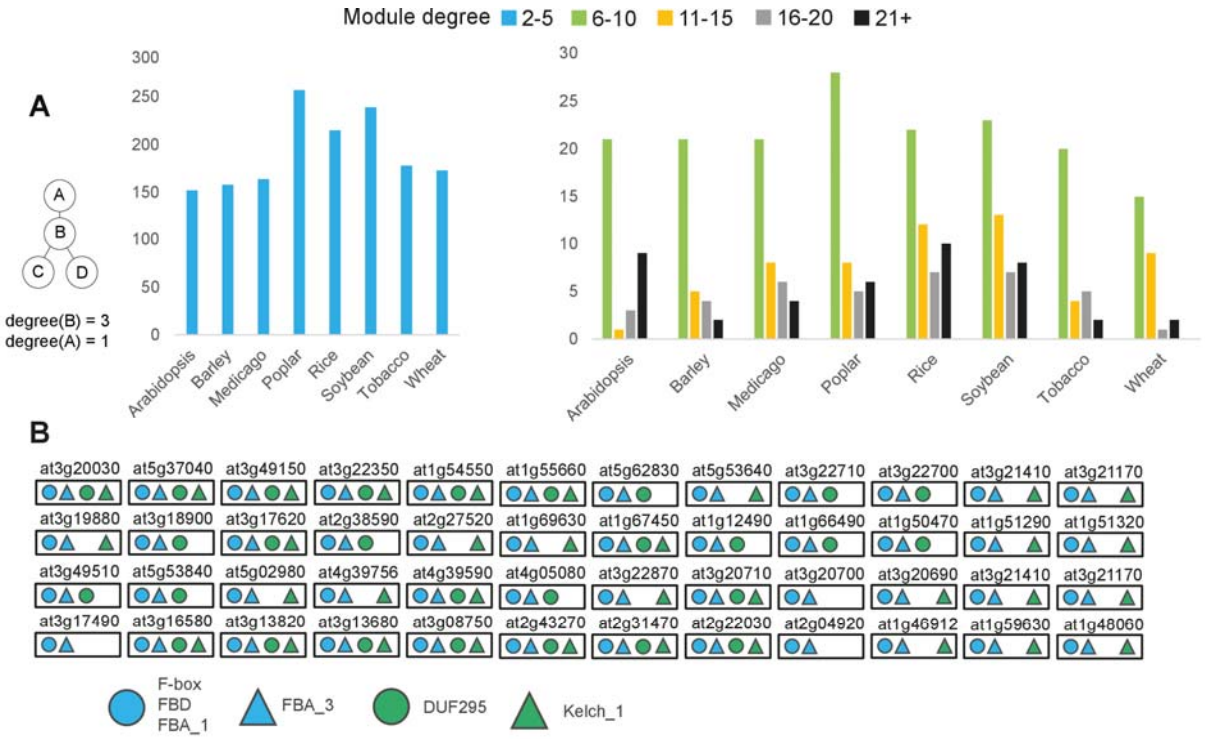

Figure 4. Distribution of module degrees. (A) Module degree is defined as the number of times a representative module has been multiplied (see example to the left of the graph). Blue bars (left column chart) indicate modules multiplied to 2 to 5 times. Green, orange, grey and black bars indicate modules with higher multiplication (right column chart). (B) Example of highly multiplied protein degradation related module from Arabidopsis. The AGI codes above each box indicate a gene used to generate the neighborhood. Colored shapes indicate the label co-occurrences shared between modules. For simplicity, gene IDs and co-expression edges are omitted.

descending order (Supplementary Fig. 4). While this heuristic favors modules with high degree, we observed that ~80% of them have been multiplied a few (< 5) times (Fig. 4A, Supplementary data 5). However, we also found modules that were multiplied more

than 20 times, for example modules related to protein degradation in Arabidopsis (Fig. 4B), metabolism in tobacco, and transcription in poplar (Supplementary Fig. 5). Taken

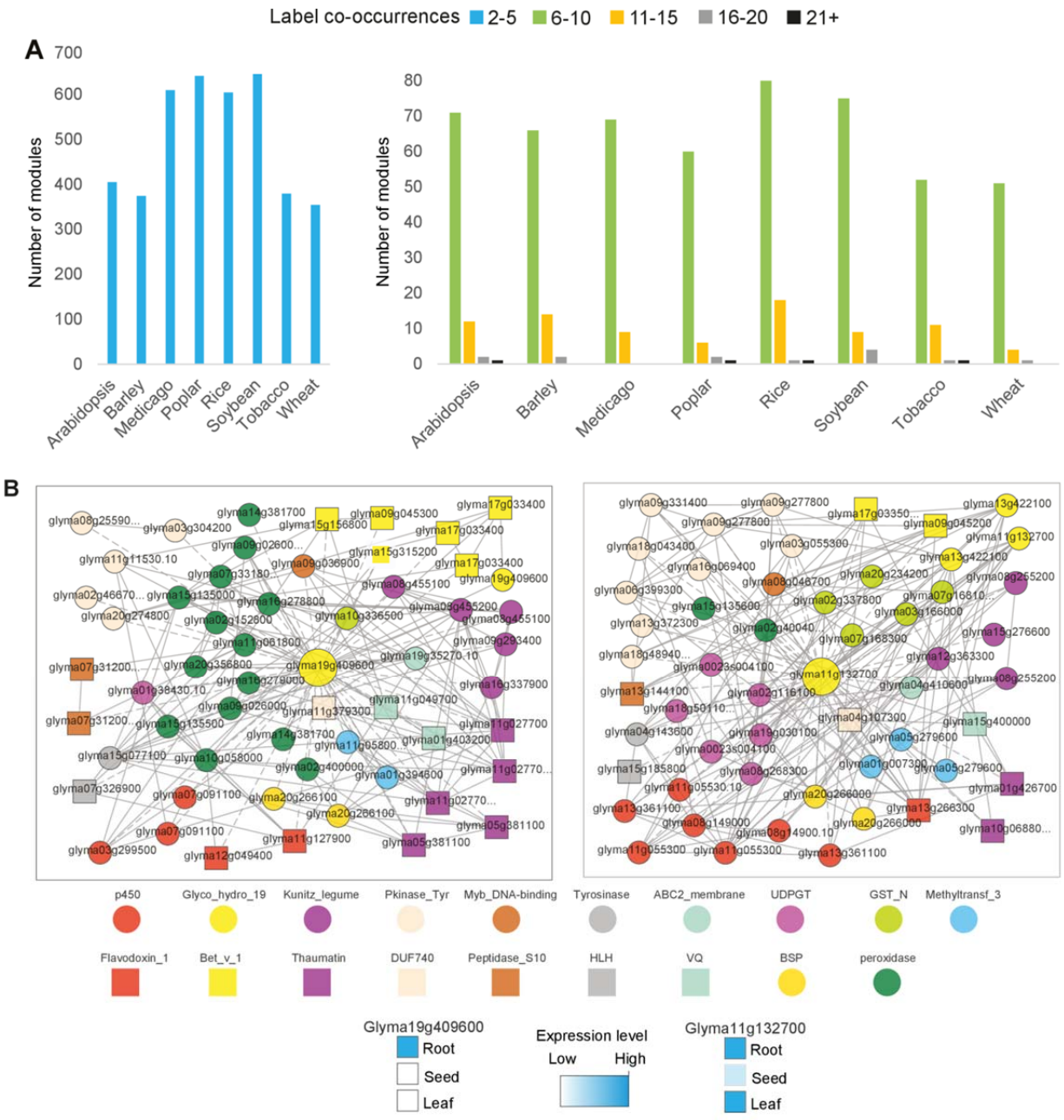

Figure 3. Distribution of label co-occurrences found between similar modules. (A) Distribution of label co-occurrences between similar modules in the eight angiosperms. Blue bars (left column chart) indicate modules similar due to 2 to 5 label co-occurrences. Green, orange, grey and black bars indicate modules similar due to higher number of label co-occurrences (right column chart). (B) Example of two modules similar due to 19 label co-occurrences in soybean, with Glyma19g40960 and Glyma11g13270 used as module centers (large yellow nodes). The colored nodes represent label co-occurrences, while grey edges represent co-expression relationships. Expression profiles of the two modules centers are found on PlaNet homepage. For simplicity, only pfam labels are shown in the legend below.

237 together, these results support frequent module multiplications, which can lead to

238 alternative pathways.

To evaluate if particular biological processes have been preferentially multiplied, we analyzed the modules by Mapman ontology enrichment analysis (Supplementary Data 6). We found that modules of high degree were enriched for regulatory processes, including: transcriptional control, RNA processing, protein degradation, and receptor kinases (Fig. 5). Furthermore, the large number of cell wall-related modules indicates that plants have evolved multiple specialized pathways to produce, remodel and degrade cell walls (Fig. 5). Interestingly, eukaryotic protein synthesis modules are also abundant, indicating that plants might employ diverse translational machineries.

The FamNet platform as a web-based tool to search for modules

To provide the research community with a platform to explore the multiplied modules we established a web-based database, coined FamNet, that is fully integrated in the PlaNet platform (Mutwil et al., 2011). We updated gene pages in PlaNet to indicate if a gene of interest participates in multiplied modules (Fig. 6A), while new label pages show the ELA network of any label of interest, and indicate multiplied gene modules in which the label is present. The FamNet database enables viewing co-expression neighborhoods, expression profiles of genes and Gene Ontology enrichment analyses of selected modules (Fig 6A). We exemplify the usefulness of the FamNet platform below using multiplied cell wall modules and secondary metabolism related modules.

Functional characterization of cell wall modules within Arabidopsis

Primary and secondary cell wall cellulose biosynthesis are multiplied modules found in higher plants (Persson et al., 2005). However, navigating to the gene page of primary cell wall multi-copper oxidase (At1g41830), suggested that Arabidopsis contain many

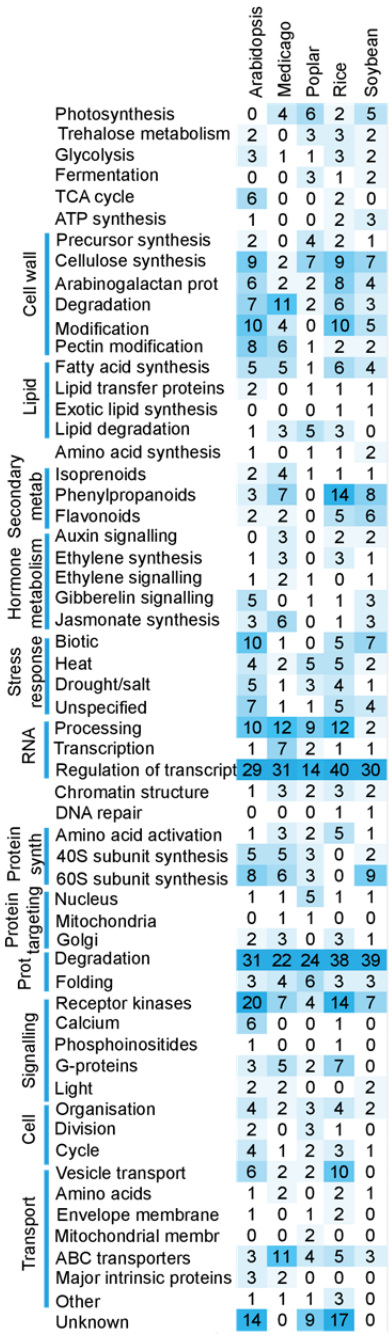

Figure 5. Gene ontology analysis of multiplied modules for the five plants with comprehensive genome sequences. The values correspond to the number of times a given ontology term was enriched in the multiplied modules.

(13) cell wall related modules similar to At1g41830 with at least 10 label co-occurrences (Fig. 6B). We chose four modules, centered around At1g41830, At5g05390, At3g13390 and At4g37160, for further analysis by selecting them from the gene module table,

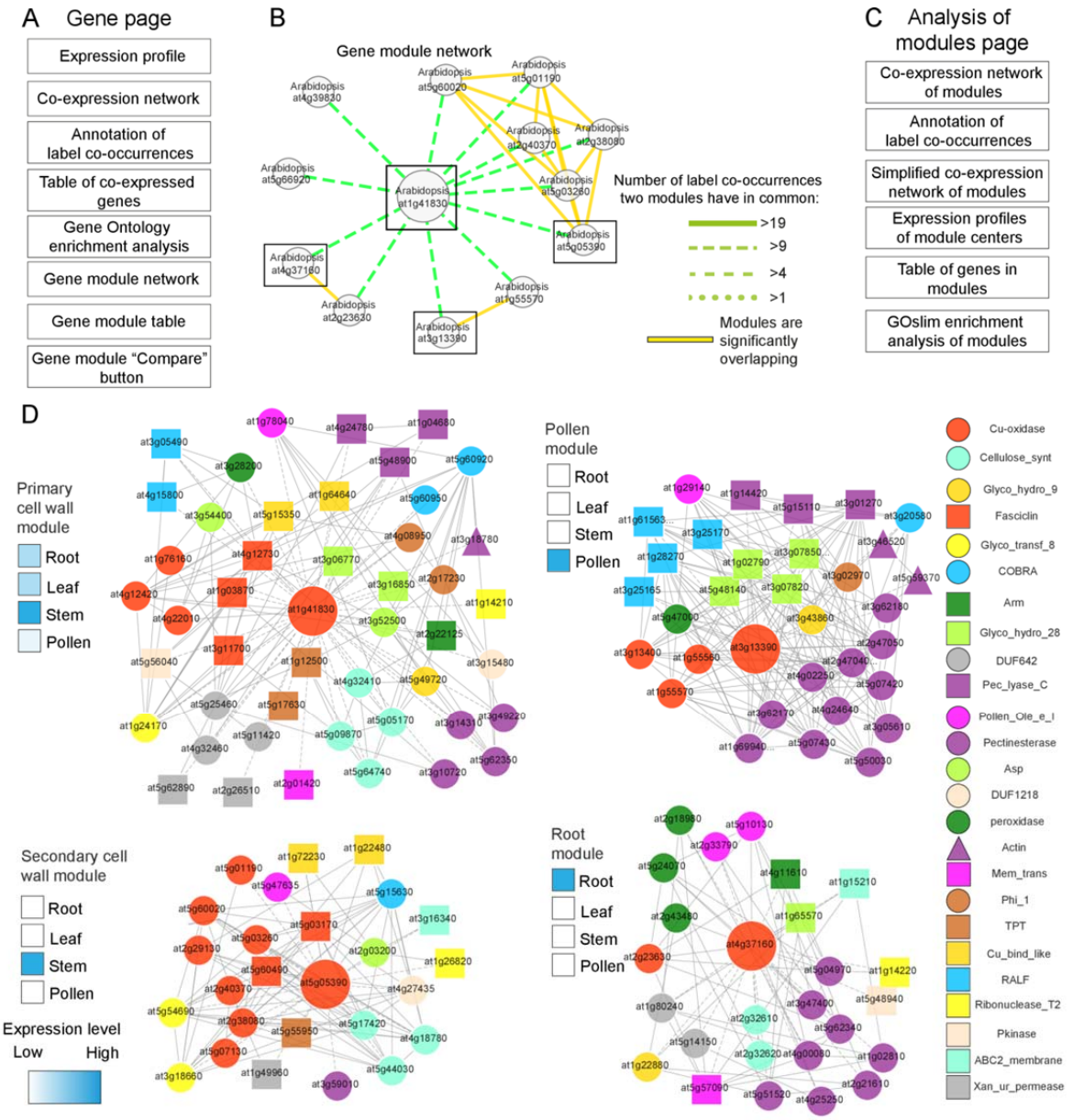

Figure 6. Cell wall biosynthetic modules occur multiple times in plants. (A) Contents of new gene pages in PlaNet. (B) Arabidopsis gene modules similar to the primary cell wall module centered around At1g41830 (large node). Green edges indicate similarity strength between modules as the number of shared label co-occurrences. The figure was generated by right-clicking on the network and selecting "toggle similarity within one species" and setting label co-occurrence cutoff of 10. Boxes indicate modules that are displayed in detail below. (C) Contents of analysis of modules page. (D) Co-expression networks of selected cell wall-related modules. Nodes and edges represent genes and co-expression relationships between genes, respectively. Colored shapes of the nodes depict label co-occurrences found in the four networks, as seen in the legend to the right. Large nodes represent genes serving as module centers. Expression profiles of module center genes were estimated from expression profiles generated by FamNet and are depicted by heat maps to the left of each module.

selecting ELA support (to remove genes not supported by ELA), and clicking Compare. Output from FamNet returned expression profile analysis of module centers (Fig. 6C), which revealed that At5g05390 is expressed in stems and co-expressed with secondary

wall-related genes. We also found that At4g37160 module contained genes preferentially expressed in roots, while At3g13390 module contained genes preferentially expressed in pollen (Fig. 6D).

To investigate the function of the pollen module further, we targeted a number of genes from this module using T-DNA insertion lines (Supplemental Data 7). Defective pollen has been reported for mutants corresponding to genes from the primary wall related module, e.g. *CELLULOSE SYNTHASE A (CESA)* genes (Persson et al., 2007), suggesting that the primary wall cellulose module is important for synthesis of the pollen wall. In contrast, none of the T-DNA mutants that corresponded to the pollen module displayed any defects in pollen morphology (Supplemental Fig. 6). However, T-DNA mutants corresponding to *COBL10*, *At4g39110*, and *At2g33420* displayed pollen tube growth-related phenotypes (Fig. 7A). COBL10 is a pollen specific homolog of COBRA, which has recently been associated with pollen tube formation (Li et al., 2013). We confirmed these results with a new T-DNA allele, *cobl10-4*, that also showed pollen tube growth defects (Fig. 7A). In contrast to the weak alleles in the previous report, *cobl10-4* showed no transmission of the T-DNA insert through pollen in reciprocal backcrosses (Fig. 7B). This phenotype could be complemented by introducing a genomic construct of *pCOBL10::COBL10* into *cobl10-4* (Fig. 7B), corroborating that COBL10 is essential for pollen tube growth. For the gene *At2g33420* we found mutant lines with bulging pollen tubes (Fig. 7A). The function of At2g33420 is unknown and based on its Pfam classification as a domain of unknown function (DUF)810, we named it CELLULOSE RELATED DUF810 (CRD)1. To confirm this *in vitro* phenotype we again performed reciprocal backcrosses which revealed that two independent T-DNA mutant lines for

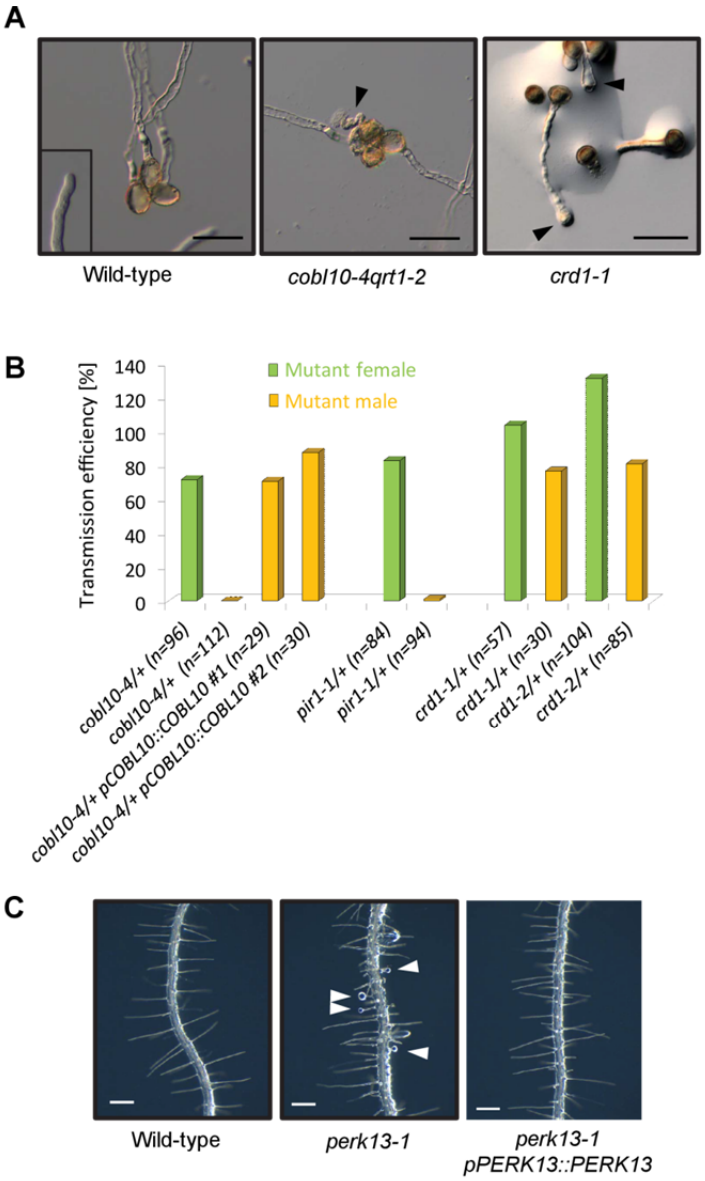

Figure 7. Pollen tube and root hair phenotypes of mutants from the pollen and root modules. (A) Pollen module mutants (cobl10-4, crd1-1) show disrupted and bulging pollen tubes (arrowheads). (B) Reciprocal backcrosses of pollen module mutants show transmission of the T-DNA insertion through male is completely abolished in cobl10-4, very strongly reduced in pir1-1, and slightly reduced in both mutant alleles for crd1. Note that the phenotype of cobl10-4 was complemented by introducing a genomic construct of COBL10. Transmission efficiency was calculated as heterozygous plants/wild-type plants*100. (C) Root module mutant (perk13-1) with bulging root hairs (arrowheads). Complementation of the perk13-1 mutant using a genomic construct of PERK13. Scale bars: 50 µm (A); 200 µm (B, D).

crd1 showed reduced transmission of the T-DNA insert through pollen (Fig. 7B). In addition, we could not obtain homozygous plants for a T-DNA mutant line corresponding to the gene *At4g39110*, and we found a segregation ratio of approximately 1:1 from a heterozygous parent plant (15 wild type: 19 heterozygous plants) suggesting

296 gametophytic defects or lethality. This gene encodes for a receptor-like kinase that is a
297 pollen specific homolog of the putative cell wall integrity sensor THESEUS (THE)1
298 (McFarlane et al., 2014). We therefore named the gene *PIRITHIOUS* (*PIR*) *1* according
299 to a friend of THESEUS in the greek mythology. To confirm pollen tube expression of
300 the gene we pollinated wild-type pistils with *pPIR::GUS* pollen (Supplemental Fig. 6).
301 Furthermore, reciprocal backcrosses showed almost no transmission of the *PIR1* T-
302 DNA insertion through the male gametophyte (Fig. 7B), which indicated pollen tube
303 growth defects.

304 Our analysis of the root specific cell wall module revealed that a T-DNA line
305 corresponding to the RLK *PERK13* displayed bulging root hair tips (Fig. 7C), which we
306 could complement by introducing a genomic *PERK13* construct into the mutant (Fig.
307 7C). These results suggested that this root specific cell wall module is associated with
308 root hair growth. Indeed, navigating to PlaNet gene page dedicated to PERK13
309 revealed enrichment for genes with annotated functions in cell wall development. To
310 conclude, the identified pollen and root modules represent specialized cell wall
311 synthesis machineries for pollen tube and root hair formation, respectively. We
312 hypothesize that these two cellulose-related modules duplicated and sub-specialized to
313 confer tip-growth in these cell types. These data indicate that our approach finds true
314 biological modules that have duplicated and attained specialized functions.

315

316 **Combining metabolomics and gene modules-Secondary metabolism**

317 Co-expression has been a rewarding approach to increase our understanding of the
318 structural pathways, and the possible regulatory machinery, governing the complexity of
319 secondary metabolism (Alejandro et al., 2012; Tohge et al., 2007). For example, this

approach has been used to find enzymes involved in distinct pathways, including steroidal glycoalkaloids (Itkin et al., 2013), flavonoid biosynthesis (Tohge and Fernie, 2010; Tohge et al., 2007; Yonekura-Sakakibara et al., 2008, 2007) as well as regulators of glucosinolate metabolism (Hirai et al., 2007) and a monolignol transporter (Alejandro et al., 2012).

Since we introduced tobacco co-expression network in the PlaNet platform, we were especially interested to try to find gene modules related to secondary metabolism in this species. To this end, we queried FamNet using several labels that might be associated with secondary metabolism. These included, “chalcone synthase”, “chalcone isomerase”, “methyltransferase_2” and “ABC transporter”. While all of these labels generated many gene modules, here we exemplify FamNet label pages by using the methyltransferase_2 label, which contain 334 genes involved in the methylation of a range of metabolites (Fig. 8A). From the resulting ELA network, it is evident that many labels that are closely related to secondary metabolism are also present in the methyltransferase_2 network, e.g. P450, transferase, peroxidase, and methyltransferase_3 (Fig. 8B). To investigate the modules that underpin the ELA network in tobacco we went to the “Network showing similar modules containing the label” section. This network shows that the ELA network is supported by many modules in all the eight plant species. In tobacco, there are 9 modules for the methyltransferase_2 ELA that are similar to one another with at least 5 label co-occurrences (Fig. 8C). While most of these modules are similar to each other (indicated by green edges) there are also several modules for which genes are showing overlapping expression patterns (yellow solid edges; Fig. 8C). To find out what genes

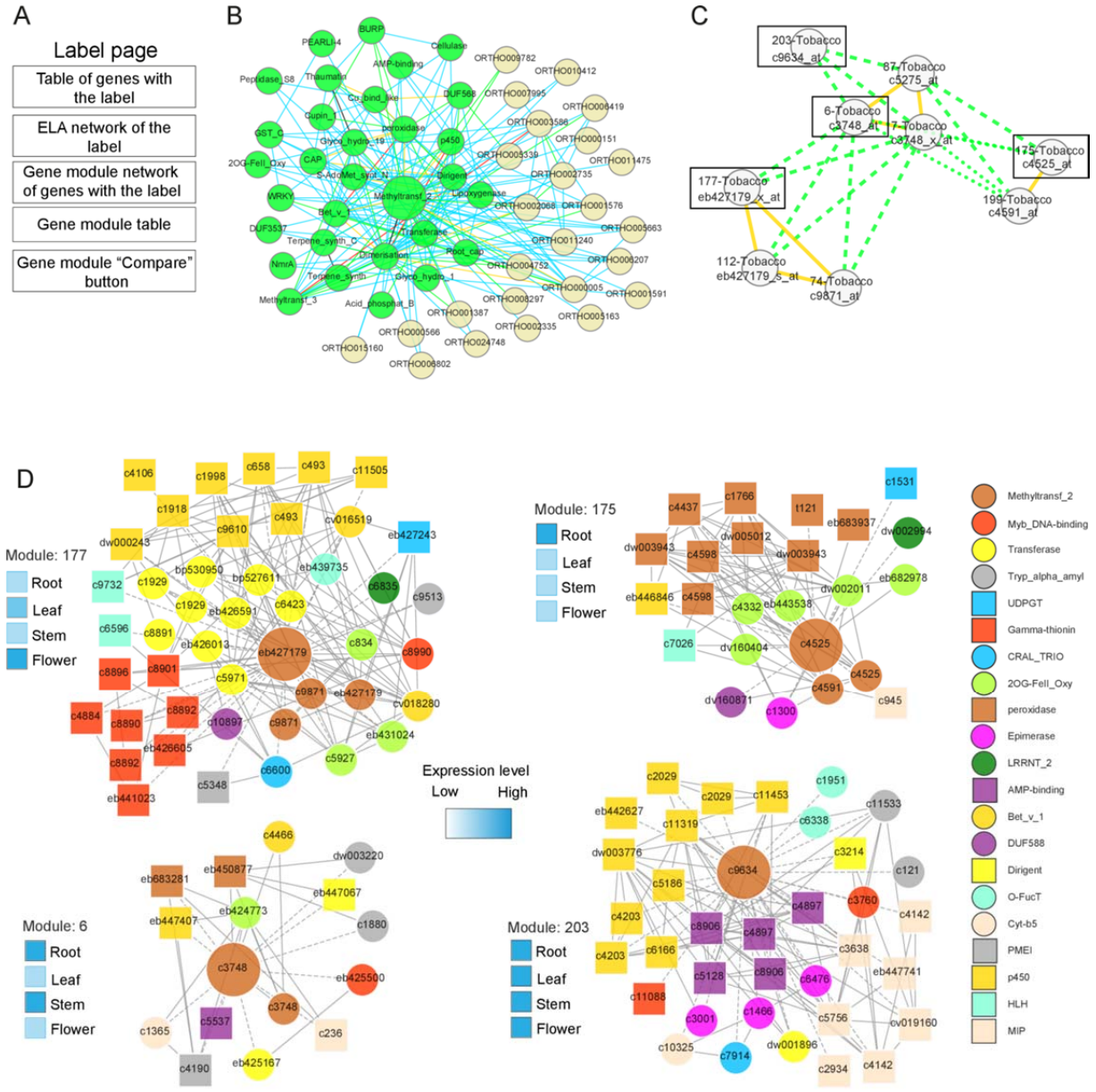

Figure 8. Secondary metabolism related modules in tobacco based on analysis of Methyltransf_2 label. (A) Contents of label pages. (B) Ensemble label association (ELA) network of Methyltransf_2. Nodes represent labels, colored edges indicate in how many species significant association was found (as in Figure 1). (C) Tobacco gene modules that contain Methyltransf_2 label. Nodes and edges are described in Figure 6. Boxes indicate modules that are displayed in detail below. Tobacco modules were highlighted by clicking on "Toggle internal similarities", and toggling all other species off. (D) Putative flavonol related modules in tobacco. Node represent genes and the colored shapes of the nodes indicate the label co-occurrence that the respective gene is associated to. Grey edges indicate co-expression relationships between the genes. Annotation of the label co-occurences is to the right. Expression profiles of module center genes are depicted by heat maps to the left of each module.

are making up these modules we selected the module for which genes did not show any overlapping expression with other modules (i.e. module 203), and one "representative" module of the modules that did show overlapping expression patterns (i.e. modules 177,

346 6, and 175; Fig. 8C). We selected these from “Table containing the modules” link,
347 selected ELA support, and clicked “Compare”. FamNet indicated that the genes being
348 the centers of these four modules have different gene expression profiles, with
349 eb427179 mainly expressed in leaf and flower tissues, c3748 in root and stem tissues,
350 c4525 in roots and c9634 ubiquitously expressed (Fig. 8D). Thus, the label
351 methyltransferase_2 is present in 9 tobacco modules that contain center genes with four
352 different expression profiles.

353 In an attempt to associate the modules with metabolite contents we first
354 performed LC-MS on plant extracts from 11 tissues, namely mature roots, young leaf,
355 mature leaf, senescent leaf, lower stem, upper stem, young silique, closed buds, open
356 buds, flower and mature seed of tobacco (*Nicotiana tabacum*) as described by (Tohge
357 and Fernie, 2010)(Supplemental Data 8 and 9). In total 105 peaks were detected by LC-
358 MS analysis, fourteen of these could be associated with three different compound
359 classes, i.e., hydroxycinnamates (chlorogenates), flavonoids (quercetin and kaempferol
360 glycosides) and diterpenes (nicotianosides), that we annotated in tobacco tissues
361 (Supplemental Fig. 7; Supplemental Data 8 and 9). Fig. 9 and Supplemental Fig. 7
362 show heatmaps and the relative relationship for the different compounds and tissues.
363 These data revealed that many compounds were preferentially accumulated in certain
364 tissues. Most of the identified compounds were present at relatively high levels in leaves
365 and in buds and flowers, but at lower levels in mature roots, mature seeds and stem
366 tissues. The amounts appeared to increase with maturity of the tissues. Whilst this
367 pattern generally holds true for the peaks detected by LC-MS, it is interesting to note
368 that a number of compounds also are exclusively present at high levels in mature roots

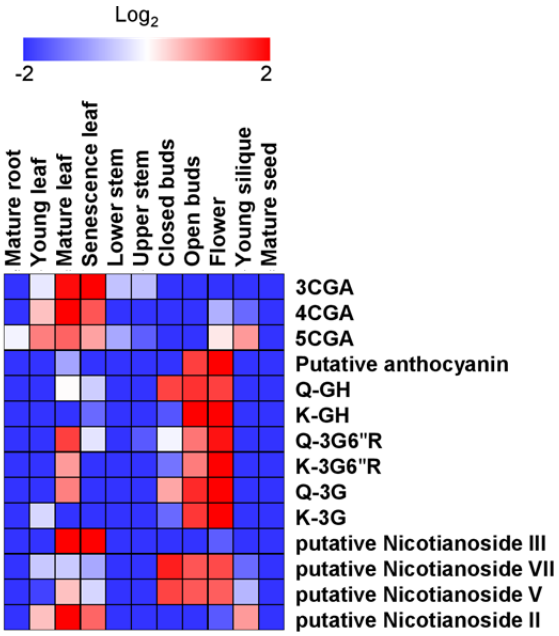

Figure 9. Heat-map visualization of secondary metabolite contents analysed by LC-MS in tobacco tissues. The analysis was conducted with three independent biological replicates. Metabolite identification and annotation were performed using standard compound literatures and co-elution profiles with tomato pericarp extracts. The relative peak area was normalized by average value and shown with logarithmic scale. Fold change is visualized by indicating color, red (high) and blue (low), respectively. Abbreviations: 3CGA, 3-caffeoylquinate; 4CGA, cryptochlorogenate; 5CGA, neochlorogenate; Q, quercetin; G. Glucose; R, rhamnose; H, hexose.

and seeds (Supplemental Fig. 7). Similar observations have previously been made for Arabidopsis (Lepiniec et al., 2006).

Then, to link the modules with metabolite profiles, we focused on a tissue in which a distinct profile of metabolites was evident. As we found many flavonols associated with floral tissues (Fig. 9, Supplemental Fig. 7), we decided that this could be an interesting and revealing example. Only genes from eb427179 cluster of overlapping modules show strong expression in tobacco flowers overlapping modules 112, 177, and 74). These overlapping modules include genes assigned to labels such as p450, Transferase and 2OG-FeII_Oxy (Fig. 8D). To get a closer estimate of the actual function of these modules, we manually investigated gene contents of the largest module 177, with gene eb427179 as center (Fig. 10). We navigated to the eb427179

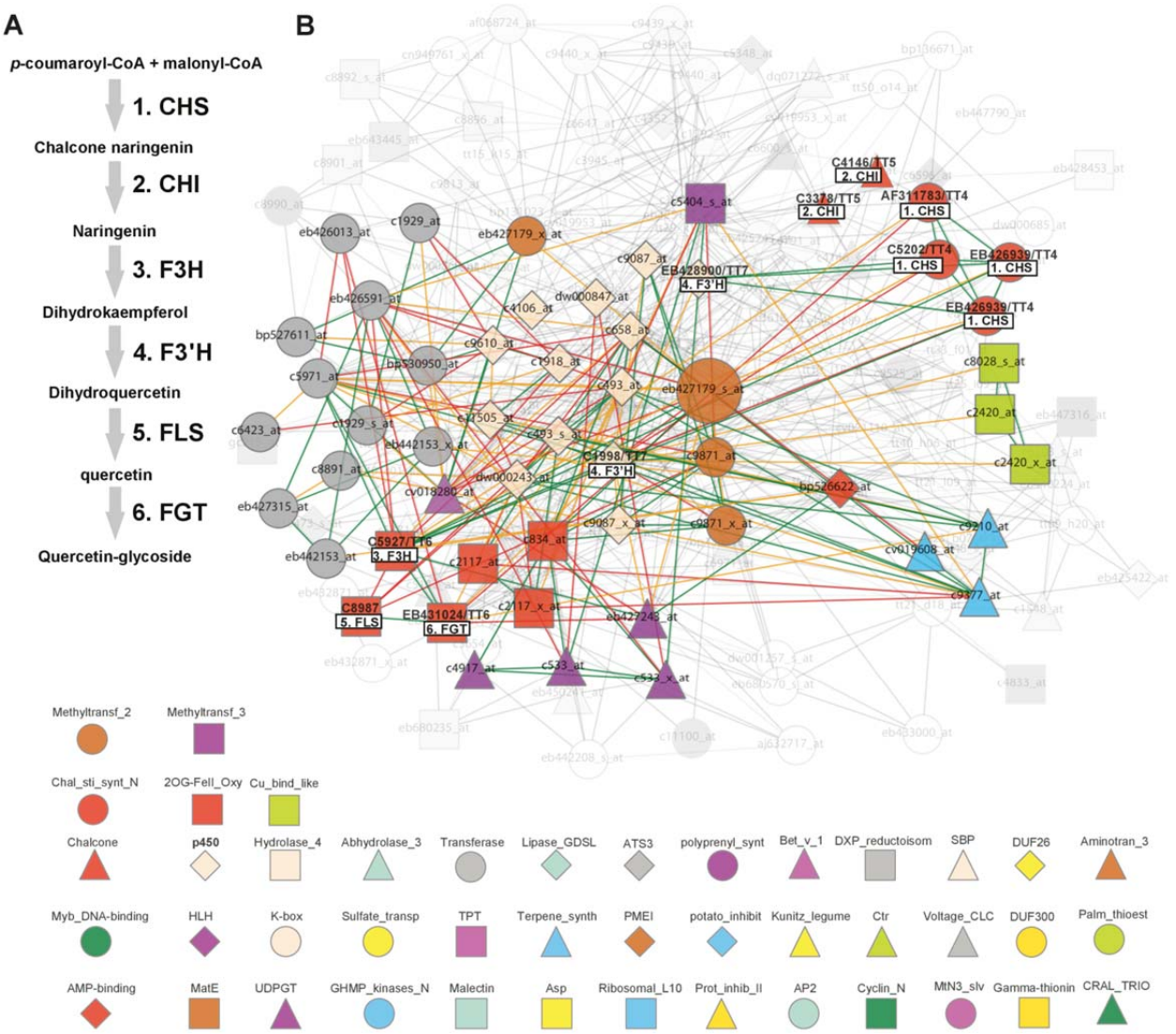

Figure 10. Scheme and co-expression network for a putative flavonol synthesis pathway in tobacco flowers. (A) Outline of a potential flavonoid synthesis pathway for tobacco (based on metabolites measured in Figure 7 and Supplemental Figure 6). CHS: CHALCONE SYNTHASE; CHI: CHALCONE ISOMERASE; F3H: FLAVANONE 3-HYDROXYLASE; FLS: FLAVONOL SYNTHASE; FGT: FLAVONOL GLYCOSYL TRANSFERASE. (B) Co-expression network of EB427179 (large node) corresponding to tobacco Methyltransf_2 Module 177 (Figure 8). Nodes are depicting genes (probeset IDs are associated with nodes), and edges depict co-expression relationships as outlined in Mutwil et al. (2011). Colored shapes of the nodes indicate the label co-occurrence that the respective gene belongs to. Genes that correspond to enzymes in the flavonoid pathway scheme (A) are highlighted in bold and are associated with respective boxes. Grayed-out, transparent part of the network represents nodes that are not supported by the ELA.

gene page by clicking on the link in the first table found on the methyltransf_2 label page. While gene ontology enrichment analysis suggests that the module is involved in terpenoid metabolic process (Supplemental Data 10), manual inspection of genes in this module revealed that 41 out of 150 genes are associated with flavonoid biosynthesis

(Supplemental Data 11). In contrast, only 4 genes could be assigned to terpenoid biosynthesis. Moreover, many of these genes encode proteins that could facilitate a direct pathway for the synthesis of the flavonoids observed in the floral tissues of tobacco (Fig. 10A and B). For example, we found all the genes corresponding to proteins that may convert 4-coumaroyl-CoA to a quercetin glycoside. These data are clearly in line with our metabolic estimates, and support the notion that the detection of modules, together with metabolic profiling, may provide a means to discover genes associated with certain metabolic processes. We hypothesize that the discrepancy between functions predicted by Gene Ontology enrichment and those derived by manual inspection of the module contents are due to incomplete/erroneous Gene Ontology annotations. Our results are further supported by looking at genes that are supported by the ELA network (Fig. 10B; function Toggle nodes supported by ELA by right-clicking on co-expression network). This function removed ~100 nodes (indicated as grayed-out, transparent nodes/edges, Fig. 10B), but retained flavonoid biosynthesis-related genes, with exception of c3378 and C4146 (chalcone isomerases). Hence, we show how ELA can be used to trim co-expression networks and to highlight conserved associations. However, this procedure might also lead to removal of relevant functions of a module as seen with the chalcone isomerases Based on these results, we suggest that the overlapping modules 112, 177 and 74, with genes preferentially expressed in flowers, represent a floral flavonoid pathway in tobacco.

To see if similar modules also are present in other dicot species, we identified the modules most similar to the floral tobacco modules in Arabidopsis. We did this by navigating to gene page of eb427179 by using probeset ID EB427179_s_at in PlaNet

(Fig. 10B). Under the heading “Gene module network” we selected modules from Arabidopsis that were linked to the *EB427179* tobacco module (Supplemental Fig. 8A, blue connections). These included modules centered around *At1g76790*, *At1g21100*, *At1g21130*, *At5g54160*, *At5g53810*, and *At5g37170* of which the latter two were overlapping modules (Supplemental Fig. 8A). Interestingly, only the overlapping modules centered around *At5g53810*, and *At5g37170,* contained genes that clearly were expressed in Arabidopsis flowers (Supplemental Fig. 8B). Closer examination revealed that these modules contained genes that were similar to the putative floral tobacco flavonol module and contained genes annotated as MYB transcription factors, cytochrome p450, methyltransferase, and UDP glucosyltransferase (Supplemental Fig. 8B). It therefore appears that tobacco and Arabidopsis both contain flavonoid-related flower-expressed modules.

While our data illustrate the power of finding commonalities within and across species for the methyltransferase_2 related modules, it is important to note that it is useful to try different centers (genes) of the modules to optimize the module content when comparing them across different species and/or within one species. This is because the co-expressed gene neighborhoods are different between homologous genes, and to obtain a complete picture of the similarities in co-expressed gene neighborhoods it is advisable to use multiple starting points for any given process, i.e. several different genes or labels. For example, in the case of secondary metabolism one could assess the ELA networks, and subsequent gene modules, for methyltransferases, chalcone synthases and glycosyltransferases and then compare the output from these to capture a broader picture of the process. These analyses may

then inform targeted reverse genetics approaches to test the predictions and thus act as powerful tools for both gene functional annotation and metabolic engineering.

How are modules multiplied?

We investigated how multiplied modules are generated. Duplication of genetic material can be divided into large-scale duplications (LSD; duplication of the whole genome or of chromosomal segments) and small-scale duplications (SGD; single gene duplications, (Maere et al., 2005)). The majority of plant species have undergone at least one, and in many cases several, LSD event(s) in form of genome duplications and/or triplications (Bowers et al., 2003). LSD events can lead to pathway multiplication in plants, as proposed for six putative modules in Arabidopsis (Blanc and Wolfe, 2004). However, multiple subsequent SGD events could also generate modules (Figure 11A). To determine whether the LSD or SGD events preferentially multiply modules, we first considered that LSD-duplicated genes can belong to three different classes in terms of modules (Figure 11B). The “across two modules” class represents LSD gene pairs found in two similar modules, and would thus support a LSD-based generation. The “within a module” and “not in module” classes represent LSD pairs found either together in one module or not in similar modules, which would reject the LSD-based generation (Figure 11B, Supplementary Figure 9). By counting the number of the three classes, we found that only 13% of LSD gene pairs were associated with ”across two modules” class, indicating that LSD events were not the predominant mechanism for module generation (Figure 11C).

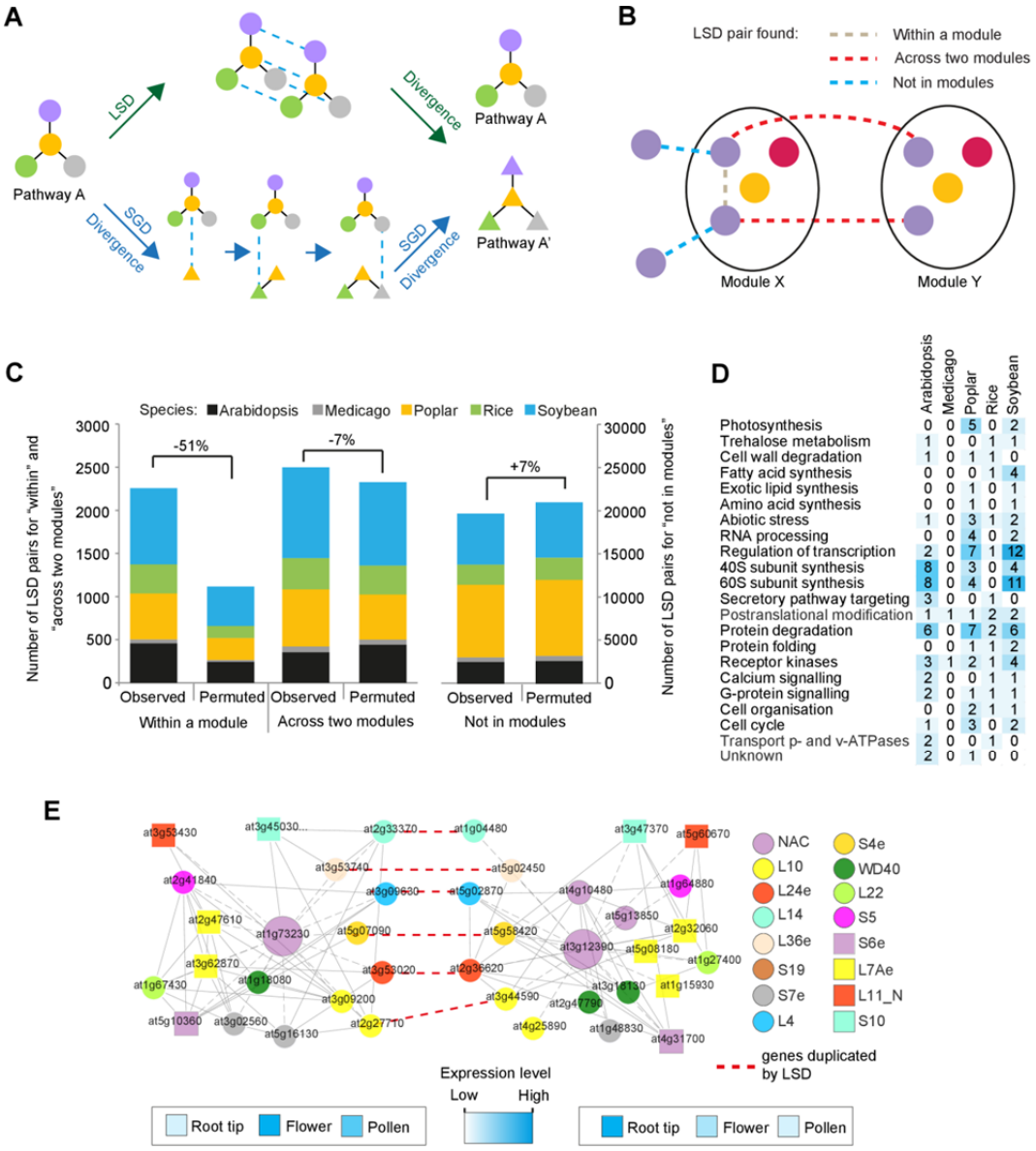

Figure 11. Gene modules are not generated through large-scale gene duplication events. (A) Two possible models for multiplying biological pathways. Colored shapes and edges represent genes and functional relationships between genes, respectively. Blue dashed edges depict recently duplicated genes. (B) Three LSD types that can occur between two similar modules. Colored shapes represent gene families. Grey, red and blue edges depict LSD-generated gene pairs that were retained within same module, found across the modules, or not found within both modules, respectively. (C) Colors and height of the bars represent species and number of LSD-generated genes. Numbers denote % change between the observed and the average of permuted networks. (D) Ontology analysis of modules significantly enriched for LSD gene pairs. (E) LSD-enriched Arabidopsis modules involved in eukaryotic ribosome biosynthesis. Colored shapes represent label co-occurrences (key shown on right panel). Grey and red dashed edges represent co-expression relationships and LSD gene pairs, respectively. Heat maps represent expression levels of the module centers, genes AT1G73230 and AT3g12390.

To further corroborate this finding, we determined the bias of the distribution of the three LSD classes by switch randomization analysis (Supplementary Figure 10). We found that the largest difference between observed and permuted networks was associated

with the “within a module” class, as the number of LSD gene pairs belonging to this class decreased by 51% (Figure 11C). This indicates that LSD-generated gene pairs tend to retain the expression profiles and thus connectivity in the co-expression networks. Conversely, the “across two modules” class decreased by only 7%, indicating that LSD gene pairs are rarely used to generate modules. Interestingly, ontology analysis of the few modules enriched for LSD gene pairs revealed that they were preferentially dedicated to biogenesis of eukaryotic ribosomes (Figures 11D-E). Taken together, the low number of LSD gene pairs in the “across modules” class, together with the modest decrease of the class in permuted networks, suggests that multiple SGD events are major contributors for the generation of modules in plants.

Conclusion

Co-expression has emerged as an important tool to rapidly infer functional relatedness among genes. These types of analyses are largely done on a gene-by-gene basis, in which a query gene for a certain biological process is used to obtain other genes that may be involved in the same process. More recently, similarities in co-expression patterns across species have become a focus (Ficklin and Feltus, 2011; Heyndrickx and Vandepoele, 2012; Langfelder and Horvath, 2008; Mutwil et al., 2011); however, instead of the gene-based approach we have here exploited the idea that sets of genes, or modules, have related functions. By analyzing such modules we constructed FamNet, which goes beyond the gene-by-gene approach to look at transcriptional associations between gene labels. The inclusion of multiple species in the FamNet platform allows for better accuracy due to conserved associations between labels. The combination between the FamNet platform and the gene-based network tool PlaNet (Mutwil et al.,

2011), will provide plant biologists with a versatile toolbox to explore conserved co-expressed relationships, which might facilitate rapid knowledge transfer within and across species.

Materials and Methods

Generation of co-expression networks for *Nicotiana tabacum*

The 144 microarrays comprising different tissues and environmental perturbation of *Nicotiana tabacum* were downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress/). The microarrays were RMA-normalized with Affymetrix Power Tools (http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx) with command line: "apt-probeset-summarize.exe -a rma -d ATCTOBa520488.cdf -o tobaccoRMA --cel-files cel_files.txt". The normalized expression values were used to generate Highest Reciprocal Rank (HRR) network with HRRnetworkCreator.py script downloaded from (http://gene2function.de/download.html, (Mutwil et al., 2011)). The co-expression networks are available at http://gene2function.de/download.html.

Assignment of pfam and PLAZA labels to genes and probesets

Fasta sequences of Pfam-A release 27 were downloaded from (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam27.0m, (Finn et al., 2006)). Protein coding sequences of genes for Arabidopsis, Medicago, poplar, rice and soybean were blasted against the PFAM-A database with e-value cut-off of 10^{-5}. For barley, wheat and tobacco, translated representative sequences, as provided by Affymetrix,

503 were used to blast against PFAM-A database (e.g.,
504 http://www.affymetrix.com/catalog/131517/AFFY/Wheat-Genome-Array#1_3). HOM and
505 ORTH gene labels for Arabidopsis, medicago, poplar, rice and soybean were
506 downloaded from PLAZA
507 (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v2_5/download/index, (Proost
508 et al., 2009)).
509
510 Co-expression networks
511 The networks, except for the Tobacco, were downloaded from PlaNet
512 (http://gene2function.de/download.html, (Mutwil et al., 2011)). The networks, together
513 with MapMan ontologies, pfam and PLAZA labels can be downloaded from PlaNet
514 (www.gene2function.de/download.html). A table summarizing properties such as
515 density, Pfam and PLAZA annotations are available as Supplementary Data 1.
516
517 Identification of gene modules
518 The pipeline is explained in detail in Supplemental material methods, which
519 consists of 2 main sections: (i) how the ELA network is generated (Section 1) and (ii)
520 how ELA is used to detect multiplied gene modules (Section 2). To generate the ELA,
521 the gene co-expression networks are first transformed into label-co-expression
522 networks (Section 1.1). Then, label associations are calculated from the label co-
523 expression networks of each species (Section 1.2). The information from the different
524 species are finally combined to generate the ELA network (Section 1.3). To detect
525 multiplied modules, first the non-conserved co-expression relationships between genes
526 are removed using the ELA (Section 2.1). Then, the similarities of neighborhoods are

estimated based on counting of label co-occurrences and the significance of neighborhood similarity is calculated by permutation analysis (Section 2.2). Finally, similar neighborhoods are summarized to arrive at gene modules (Section 2.3), and overlapping modules are detected (Section 2.4).

Estimating distribution of label co-occurrences between gene modules

To obtain a global collection of label co-occurrences from non-overlapping gene modules we used a greedy heuristic, which sorted similar duplicated module pairs according to the number of label co-occurrences, in decreasing order (Supplementary Fig. 1). If at least one of the modules contains genes that have not been collected before, the value of the shared label co-occurrence in the module-pair is collected. The heuristic (i) sorts all module-pairs according to their label co-occurrence value. (ii) For each module-pair, each module is compared to the “takenGenes” set. (iii) If genes from one or two of the modules are not in “takenGenes”, the label co-occurrence value is collected, and genes from either one or both modules are added to the “takenGenes” set. (iv) Steps (ii-iii) are repeated for each module pair obtained in step (i). The heuristic is exemplified in Supplementary Fig. 1 and the result, which was used to generate Fig. 3A, is shown on Supplementary Data 1.

Estimating distribution of module degrees

We used a greedy heuristic to estimate the degree, i.e., the copy number of gene modules (Supplementary Fig. 4). Similar to the previous section, the heuristic (i) sorts all module centers in descending order according to the number of other modules they are similar to. (ii) Content of each module center is compared to “takenGenes” set. (iii) If

genes from a module center are not in “takenGenes”, the module degree value is collected, and genes from the module center, together with genes from similar modules are added to the “takenGenes” set. (iv) Steps (ii-iii) are repeated for each module center obtained in step (i). The heuristic is exemplified in Supplementary Fig. 4 and is used to generate Fig. 4A in the main text. Similar to the estimation of distribution of similarity strength between gene modules, the greedy heuristic returns a lower bound of actual number of modules. The results used to generate Fig. 4A are shown in Supplementary Data 5.

Estimating functional ontologies of module-pairs

We used MapMan ontologies to investigate functional enrichment of multiplied modules (Klie and Nikoloski, 2012). The empirical p-value of ontological term enrichment is conducted by first estimating the number of ontologies present in each module, followed by shuffling gene-ontology assignments 1000 times. The empirical p-value is given by the proportion of scores from the shuffling which are larger than the score from the original network. Finally, the analysis estimates which enriched MapMan terms are shared between two modules, and assigns shared ontologies to the modules. The results from this analysis can be found in Supplementary Data 6. Fig. 5 was generated by counting ontology terms of duplicated modules, where module selection is the same as used to generate Fig. 3A (see above). To emphasize more complex modules, we have used label co-occurrence cutoff of 5, i.e., two modules share at least 5 label co-occurrences. Number of enriched ontologies for cutoffs of 2, 5 and 10 can be found in Supplementary Data 6.

Plant material, growth conditions and mutant analysis

Seeds for all mutant lines were obtained from the Nottingham Arabidopsis Stock Centre (NASC, http://arabidopsis.info) and are all in Col-0 background (Supplementary Table 7). Primers for genotyping are listed in Supplementary Table 5. Mutants from the pollen module were first grown on MS medium containing 1 % sucrose for 2 weeks, and then transferred to standard soil (Einheitserde GS90; Gebrüder Patzer, Sinntal-Jossa, Germany) and grown in a greenhouse under a 16 h light/ 8 h dark regime at 21°C (day) and 17°C (night). Pollen tube growth assays were performed as previously described by (Boavida and McCormick, 2007). Observations of pollen tubes were carried out with a BX61 microscope (Olympus, Hamburg, Germany) equipped with differential interference contrast microscopy, using a 10x objective. Imaging was carried out with a ColorviewIII digital camera (Olympus, Hamburg, Germany) controlled with the cell^P software from Olympus. Mutants from the root module were grown on MS medium containing 120 mM sucrose for 10 days under long day conditions (16 h light/ 8 h dark). Note that the phenotype of the *perk13* mutant is conditional on 120 mM sucrose.

Metabolite profiling and data analysis

Secondary metabolite analysis by LC-MS was performed as described by (Tohge and Fernie, 2010). Obtained chromatograph data was processed using Xcalibur 2.1 software (Thermo Fisher Scientific, Waltham, USA). The obtained peak matrix was normalized using the internal standard (isovitexin, CASRN: 29702-25-8). Metabolite identification and annotation were performed using standard compounds (3-caffeoylquinate, 3CQA; rutin, Q-3G6"R; kaempferol-3-*O*-rutinoside, K-3G6"R; quercetin-3-*O*-glucoside, Q-3G; kaempferol-3-*O*-glucoside, K-3G), spectral data described in

599 literatures (Bedoya et al., 2012; Heiling et al., 2010; Jassbi et al., 2008; Niggeweg et al.,
600 2004; Onkokesung et al., 2012) and co-elution profile with tomato pericarp extracts
601 (Rohrmann et al., 2011).
602
603 Estimating types of large-scale duplicated genes (LSD) in modules
604 We have used the PGDD database to retrieve genes duplicated by large-scale
605 duplications for each of the five sequenced species
606 (http://chibba.agtec.uga.edu/duplication/, (Lee et al., 2013)). The large-scale
607 duplications (LSD) encompass genome and chromosome segment duplications, and
608 contain gene-pairs that were found to be generated by LSD. We have defined 3 types of
609 relationships LSD gene pairs can have: (i) both of the two LSD genes are found in two
610 similar modules, (ii) both genes are found in same module and (iii) the LSD gene pairs
611 cannot be assigned to type (i) and (ii). It is important to note that a gene pair can be
612 present in multiple modules, and can therefore have multiple LSD relationships
613 (Supplementary Figure 9). Here, we have set the order of relationships to (i)>(ii)>(iii).
614 For example, if a LSD gene pair is determined to be both within a module and across
615 two modules (such as genes 2 and 4, Supplementary Figure 9), the analysis assigns
616 the LSD pair to “within module” relationship. The rationale behind setting this order is
617 twofold: First, since LSD relationships are investigated for each module pair, and the
618 maximum number of genes in a module usually does not exceed 50, the majority of
619 LSD gene-pairs are always assigned to type (iii) for each module pair comparison.
620 Consequently, if a relationship (i) or (ii) is detected, it has higher precedence over
621 relationship (iii). Second, relationship (i) represents an LSD gene pair that is co-

622 expressed to some degree (genes 2 and 4 are connected via gene 3, Supplementary
623 Figure 9) Hence, there is an uncertainty whether the gene pair is (a) part of the same
624 module (module C), or (b) two similar modules with very close expression profiles
625 (module A and E). Here, we choose (i)>(ii), to select the more conservative scenario.
626 Using this strategy, we have counted the number of the three LSD types for the five
627 sequenced plant species. The outcome of this analysis is shown on Figure 11.
628
629 Switch randomization of LSD types to determine significance of LSD types distribution
630 In this section we aim to investigate if there is a bias in distribution of the “within”,
631 “across” and “not in modules” edges described in the previous section. Permuting the
632 LSD edges should indicate if LSD gene-pairs are preferentially found in the “within”,
633 “across” and “not in modules”. To do this, we have employed switch randomization
634 analysis of the LSD edges with two constraints: (i) LSD genes must belong to same
635 family and (ii) edges have to be shuffled to other members of the family (Supplementary
636 Figure 10). The permutation analysis was repeated 1000 times and the number of each
637 of the “within”, “across” and “not in modules” relationships was noted for each
638 permutation (Supplementary Figure 10B). The average of the analysis was used to
639 generate the “permuted data” bars on Figure 11C.
640
641 **Figure legends**
642 **Figure 1. Generating Ensemble Label Association (ELA) network.** (A) Co-
643 expression networks derived from the PlaNet platform are used as input. Each gene
644 may be assigned multiple labels. (B) The gene co-expression networks are
645 decomposed into label co-expression networks, where nodes represent labels assigned

to genes, and edges represent co-expression relationships between the labels. (C) Associations between labels are detected for each species by label-based node cover and permutation test. The result is label association networks, where nodes represent labels and edges represent associations between the labels. (D) Label association networks are combined into ensemble label association (ELA) network. The number of edges (associations) that are conserved across the different species is determined. In this example, labels C and D are connected in three species (species 2, 3 and 4). (E) ELA of Cu-oxidase_2. (F) ELA of PSI_PSAK. (G) ELA of NAC. Green and yellow nodes represent Pfam and PLAZA labels, respectively. Edges show in how many species an association was found.

Figure 2. Detecting similar modules. The pipeline is exemplified by searching for similar modules to the neighborhood of gene 1. (A) The neighborhood of the query gene 1 is first isolated. Nodes represent genes, edges represent co-expression relationships and node colors indicate labels found in collected genes. Note that gene 2 has two labels, red and blue. Label neighborhoods of genes containing orange label (genes 6 and 11) are isolated. (B) The neighborhoods are trimmed with ELA, where labels not supported by ELA are removed (Fig. 1D). (C) Label co-occurrences found between the neighborhood of the query gene and label neighborhoods are calculated. As gene 2 contains two labels, genes 7 and 8 are collapsed into one label co-occurrence. (D) The significance of found label co-occurrences is estimated by permutation analysis. Green edges indicate similar neighborhoods. (E) Overlapping modules are identified. (F) Total number and percentage of genes assigned to similar modules. Blue line (left y-axis) denote number of genes assigned to modules. Gray bars

(right y-axis) represent the percentage of total genes found on the microarrays that are assigned to modules (right y-axis). Note that % of genes for barley, wheat and tobacco is missing due to the lack of comprehensive genome annotation of the microarrays.

Figure 3. Distribution of label co-occurrences found between similar modules. (A) Distribution of label co-occurrences between similar modules in the eight angiosperms. Blue bars (left column chart) indicate modules similar due to 2 to 5 label co-occurrences. Green, orange, grey and black bars indicate modules similar due to higher number of label co-occurrences (right column chart). (B) Example of two modules similar due to 19 label co-occurrences in soybean, with Glyma19g40960 and Glyma11g13270 used as module centers (large yellow nodes). The colored nodes represent label co-occurrences, while grey edges represent co-expression relationships. Expression profiles of the two modules centers are found on PlaNet homepage. For simplicity, only pfam labels are shown in the legend below.

Figure 4. Distribution of module degrees. (A) Module degree is defined as the number of times a representative module has been multiplied (see example to the left of the graph). Blue bars (left column chart) indicate modules multiplied to 2 to 5 times. Green, orange, grey and black bars indicate modules with higher multiplication (right column chart). (B) Example of highly multiplied protein degradation related module from Arabidopsis. The AGI codes above each box indicate a gene used to generate the neighborhood. Colored shapes indicate the label co-occurrences shared between modules. For simplicity, gene IDs and co-expression edges are omitted.

694 **Figure 5. Gene ontology analysis of multiplied modules for the five plants with**
695 **comprehensive genome sequences**. The values correspond to the number of times a
696 given ontology term was enriched in the multiplied modules.
697
698 **Figure 6. Cell wall biosynthetic modules occur multiple times in plants.** (A)
699 Contents of new gene pages in PlaNet. (B) Arabidopsis gene modules similar to the
700 primary cell wall module centered around At1g41830 (large node). Green edges
701 indicate similarity strength between modules as the number of shared label co-
702 occurrences. The figure was generated by right-clicking on the network and selecting
703 “toggle similarity within one species” and setting label co-occurrence cutoff of 10. Boxes
704 indicate modules that are displayed in detail below. (C) Contents of analysis of modules
705 page. (D) Co-expression networks of selected cell wall-related modules. Nodes and
706 edges represent genes and co-expression relationships between genes, respectively.
707 Colored shapes of the nodes depict label co-occurrences found in the four networks, as
708 seen in the legend to the right. Large nodes represent genes serving as module
709 centers. Expression profiles of module center genes were estimated from expression
710 profiles generated by FamNet and are depicted by heat maps to the left of each module.
711
712 **Figure 7. Pollen tube and root hair phenotypes of mutants from the pollen and**
713 **root modules.** (A) Pollen module mutants (*cobl10-4*, *crd1-1*) show disrupted and
714 bulging pollen tubes (arrowheads). (B) Reciprocal backcrosses of pollen module
715 mutants show transmission of the T-DNA insertion through male is completely abolished
716 in *cobl10-4*, strongly reduced in *pir1-1*, and slightly reduced in both mutant alleles for
717 *crd1*. Note that the phenotype of *cobl10-4* was complemented by introducing a genomic

718 construct of *COBL10*. Transmission efficiency was calculated as heterozygous
719 plants/wild-type plants*100. (C) Root module mutant (*perk13-1*) with bulging root hairs
720 (arrowheads). Complementation of the *perk13-1* mutant using a genomic construct of
721 *PERK13*. Scale bars: 50 μm (A); 200 μm (B, D).
722
723 **Figure 8. Secondary metabolism related modules in tobacco based on analysis of**
724 **Methyltransf_2 label.** (A) Contents of label pages. (B) Ensemble label association
725 (ELA) network of Methyltransf_2. Nodes represent labels, colored edges indicate in how
726 many species an association was found (as in Fig. 1). (C) Tobacco gene modules that
727 contain Methyltransf_2 label. Nodes and edges are described in Fig. 6. Boxes indicate
728 modules that are displayed in detail below. Tobacco modules were highlighted by
729 clicking on “Toggle internal similarities”, and toggling all other species off. (D) Putative
730 flavonol related modules in tobacco. Node represent genes and the colored shapes of
731 the nodes indicate the label co-occurrence that the respective gene is associated to.
732 Grey edges indicate co-expression relationships between the genes. Annotation of the
733 label co-occurences is to the right. Expression profiles of module center genes are
734 depicted by heat maps to the left of each module.
735
736 **Figure 9. Heat-map visualization of secondary metabolite contents analysed by**
737 **LC-MS in tobacco tissues.** The analysis was conducted with three independent
738 biological replicates. Metabolite identification and annotation were performed using
739 standard compound literatures and co-elution profiles with tomato pericarp extracts. The
740 relative peak area was normalized by average value and shown with logarithmic scale.
741 Fold change is visualized by indicating color, red (high) and blue (low), respectively.

742 Abbreviations: 3CGA, 3-caffeoylquinate; 4CGA, cryptochlorogenate; 5CGA,
743 neochlorogenate; Q, quercetin; G. Glucose; R, rhamnose; H, hexose.
744
745 **Figure 10. Scheme and co-expression network for a putative flavonol synthesis**
746 **pathway in tobacco flowers**. (A) Outline of a potential flavonoid synthesis pathway for
747 tobacco (based on metabolites measured in Fig. 7 and Supplemental Fig. 6). CHS:
748 CHALCONE SYNTHASE; CHI: CHALCONE ISOMERASE; F3H: FLAVANONE 3-
749 HYDROXYLASE; FLS: FLAVONOL SYNTHASE; FGT: FLAVONOL GLYCOSYL
750 TRANSFERASE. (B) Co-expression network of EB427179 (large node) corresponding
751 to tobacco Methyltransf_2 Module 177 (Fig. 8). Nodes are depicting genes (probeset
752 IDs are associated with nodes), and edges depict co-expression relationships as
753 outlined in (Mutwil et al., 2011). Colored shapes of the nodes indicate the label co-
754 occurrence that the respective gene belongs to. Genes that correspond to enzymes in
755 the flavonoid pathway scheme (A) are highlighted in bold and are associated with
756 respective boxes. Grayed-out, transparent part of the network represents nodes that are
757 not supported by the ELA.
758
759 **Figure 11. Gene modules are not likely generated through large-scale gene**
760 **duplication events.** (A) Two possible models for multiplying biological pathways.
761 Colored shapes and edges represent genes and functional relationships between
762 genes, respectively. Blue dashed edges depict recently duplicated genes. (B) Three
763 LSD types that can occur between two similar modules. Colored shapes represent gene
764 families. Grey, red and blue edges depict LSD-generated gene pairs that were retained
765 within same module, found across the modules, or not found within both modules,

766 respectively. (C) Colors and height of the bars represent species and number of LSD-
767 generated genes. Numbers denote % change between the observed and the average of
768 permuted networks. (D) Ontology analysis of modules significantly enriched for LSD
769 gene pairs. (E) LSD-enriched Arabidopsis modules involved in eukaryotic ribosome
770 biosynthesis. Colored shapes represent label co-occurrences (key shown on right
771 panel). Grey and red dashed edges represent co-expression relationships and LSD
772 gene pairs, respectively. Heat maps represent expression levels of the module centers,
773 genes AT1G73230 and AT3g12390.
774
775 **Supplementary Figure 1.** Estimating genome-wide distribution of label co-occurrences
776 between gene modules. (A) Ellipses represent gene modules, while green edges depict
777 significantly similar gene modules. Number of label co-occurrences between the
778 modules are indicated by edge styles. Overlapping ellipses indicate which modules are
779 sharing genes, i.e. overlapping. (B) Module-pair collection is sorted according to the
780 number of label co-occurrences, with more similar module-pairs being collected first. (C)
781 Here, the heuristic is determining overlapping modules, with modules having more label
782 co-occurrences having higher precedence over modules with less number of label co-
783 occurrences. For example, both modules in module-pair 2, and one module in module
784 pair 6 are overlapping with modules that have higher precedence. If at least one of the
785 modules is not overlapping with modules of higher precedence, the label co-occurrence
786 value is collected. (D) In this example, out of six module pairs, 5 label co-occurrence
787 values are collected. Note that the label co-occurrence value from pair 2 is disregarded,
788 as both modules are overlapping with pair 1.

Supplementary Figure 2. An example of large gene modules involved in chromatin remodeling in rice. (A) Two gene modules from rice with loc_os10g31970 and loc_os02g46450 used as module centers (large nodes). The nodes represent label co-occurrences, while node labels represent genes assigned to the label co-occurrences. Gray edges represent associations of the label co-occurrences to the module centers. The two modules are overlapping to some degree and consequently share genes, shown by red dashed edges. The number of dashed edges is equal to the number of genes shared between the label co-occurrences. (B) Labels found in the label co-occurrences. For simplicity, only pfam labels are shown. The two modules show enrichment in ontologies representing transcription factors, chromatin remodeling/structure factors, signaling and cell division. The ontology analysis for both modules can be viewed at http://aranet.mpimpgolm.mpg.de/responder.py?name=gene!osa!13835 and http://aranet.mpimpgolm.mpg.de/responder.py?name=gene!osa!8427

Supplementary Figure 3. An example of large gene modules involved in ribosome biosynthesis in tobacco. (A) To make comparisons of gene module content easier, the co-expression networks are simplified by collecting all genes belonging to one label co-occurrence and representing it as one node. In this example, genes B and C belong to same label co-occurrence (green node) and are assigned to the same node in simplified network. (B) Two gene modules from tobacco with C1368 and C1349 used as module centers (large nodes). The nodes represent label co-occurrences, while node labels

represent genes assigned to the label co-occurrences. Gray edges represent associations of the label co-occurrences to the module centers. The two modules are weakly overlapping and consequently sharing genes, which is shown by connecting the overlapping label co-occurrences by red dashed edges. (C) Labels found in the label co-occurrences. For simplicity, only pfam labels are shown. The two modules show enrichment in ontologies representing ribosome structural components.

Supplementary Figure 4. Estimating the distribution of representative module degrees. (A) Nodes represent modules, and edges indicate similar modules. Numbers adjacent to a module indicate the degree (d) of a module. (B) Module collection is determined by module degree, with modules with higher degree having higher precedence. (C) The first module, with highest degree is collected (module D), together with its neighbors (modules B, C,E and F). Modules can only be collected once. In this example, out of six modules, two module degrees were collected (d=2, d=4 for modules A and D).

Supplementary Figure 5. Examples of frequently multiplied modules in plants. Genes/probesets that were used as module centers are indicated above the boxes. Colored shapes indicate label co-occurrences that were present in the respective modules. For simplicity, only pfam labels are shown. (A) Metabolism related modules in barley. (B) Transcription related modules in soybean.

Supplementary Figure 6. Mutants from the pollen cell wall module show normal pollen.

834 (A-D) Whole anthers and mature pollen (inset upper right) stained with Alexander stain
835 and DAPI (insets lower left) indicate that pollen viability is not affected in the mutants.
836 Note that cobl10-4 was crossed into the quartet
837 (qrt)1-2 background, which displays tetrads of pollen grains after meiosis (Francis et al.,
838 2006). (E) Pollination of wild type pistils with pPIR::GUS pollen shows pollen and pollen
839 tube specific expression of PIR1. Scale bars: 50 μm (including insets).
840
841 **Supplemental Figure 7.** Hierarchical clustering analysis of LC-MS metabolite profile of
842 tobacco tissues. Relative peak area was normalized by average value and shown with
843 logarithmic scale (log2). Fold change is visualized by indicating color, red (high) and
844 blue (low), respectively.
845
846 **Supplementary Figure 8.** EB427179-like gene modules in Arabidopsis. A) Gene
847 module network of EB427179 with Arabidopsis modules shown. B) Expression profile of
848 At5g53810. C) Gene module comparison of EB427179 and At5g53810 and At5g37170.
849
850 **Supplementary Figure 9.** Genes can be present in multiple modules and have multiple
851 LSD relationships. Nodes represent genes, while black solid edges represent co-
852 expression relationships. Node colors represent different gene labels. Dashed edges
853 represent the three LSD relationships. In this example, genes 2 and 4 can be in the
854 same module (module C), or in two similar modules (module A and E), depending on
855 the investigated module.
856

Supplementary Figure 10. Counting and estimating the significance of large-scale duplicated genes (LSD) in modules. (A) Consider two similar modules, X and Y, containing four genes each. Nodes and node colors represent genes and labels, respectively. Red edges represent LSD pairs found across the two modules. Gray edges represent LSD pairs found within a module, while blue edges represent LSD pairs not found in two similar modules. In this example, 2 red edges, 1 gray edge and 2 blue edges (5 edges in total) were found. Note that for simplicity, only the violet label is analyzed in this example. (B) To estimate the significance of the edge distributions, the 5 LSD edges are distributed randomly among the members of the violet label. The criteria are: the number of edges must stay constant (i.e. 5) and the edges can be only distributed among the violet label. The LSD edges are permuted 1000 times, and the number of red, gray

and blue edges is counted for each permutation. The analysis is done for each label, if any LSD gene-pairs are found for the label.

Supplemental Data 1. Properties of the microarray data and co-expression networks.

Supplemental Data 2. ELA network.

Supplemental Data 3. Multiplied modules.

Supplemental Data 4. Distribution of similarity strength values (in label co-occurrences) between modules.

Supplemental Data 5. Degree vs. number of modules in the eight analyzed species.

Supplemental Data 6. MapMan ontology terms enriched between multiplied modules.

879 **Supplemental Data 7.** T-DNA insertion information about the selected genes from the
880 pollen and root specific cell wall modules.
881 **Supplemental Data 8.** Metabolite Reporting Guidelines (Checklist)
882 **Supplemental Data 9.** Recommendations for GC- and LC-MS.
883 **Supplemental Data 10.** GO analysis of EB427179_s_at
884 **Supplemental Data 11.** Functional annotation of module eb427179_s_at
885 **Supplemental Methods.** Description of algorithms used in FamNet database.
886
887

Parsed Citations

Alejandro, S., Lee, Y., Tohge, T., Sudre, D., Osorio, S., Park, J., Bovet, L., Lee, Y., Geldner, N., Fernie, A.R., Martinoia, E., 2012. AtABCG29 is a monolignol transporter involved in lignin biosynthesis. Curr. Biol. 22, 1207-1212. doi:10.1016/j.cub.2012.04.064
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Aoki, K., Ogata, Y., Shibata, D., 2007. Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol. doi:10.1093/pcp/pcm013
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Bedoya, L.C., Martínez, F., Orzáez, D., Daròs, J.-A., 2012. Visual tracking of plant virus infection and movement using a reporter MYB transcription factor that activates anthocyanin biosynthesis. Plant Physiol. 158, 1130-8. doi:10.1104/pp.111.192922
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Bergmann, S., Ihmels, J., Barkai, N., 2004. Similarities and differences in genome-wide expression data of six organisms. PLoS Biol. 2, E9. doi:10.1371/journal.pbio.0020009
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Blanc, G., Wolfe, K.H., 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16, 1667-1678. doi:10.1105/tpc.021345
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Boavida, L.C., McCormick, S., 2007. Temperature as a determinant factor for increased and reproducible in vitro pollen germination in Arabidopsis thaliana. Plant J. 52, 570-582. doi:10.1111/j.1365-313X.2007.03248.x
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Bowers, J.E., Chapman, B.A., Rong, J.K., Paterson, A.H., 2003. Unravelling Angiosperm Genome Evolution by Phylogenetic Analysis of Chromosomal Duplication Events. Nature 422, 433-438. doi:10.1038/nature01521
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Brown, D.M., Zeef, L.A.H., Ellis, J., Goodacre, R., Turner, S.R., 2005. Identification of novel genes in Arabidopsis involved in secondary cell wall formation using expression profiling and reverse genetics. Plant Cell 17, 2281-95. doi:10.1105/tpc.105.031542
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Conant, G.C., Wolfe, K.H., 2006. Functional partitioning of yeast co-expression networks after genome duplication. PLoS Biol. 4, e109. doi:10.1371/journal.pbio.0040109
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ehlting, J., Sauveplane, V., Olry, A., Ginglinger, J.-F., Provart, N.J., Werck-Reichhart, D., 2008. An extensive (co-)expression analysis tool for the cytochrome P450 superfamily in Arabidopsis thaliana. BMC Plant Biol. 8, 47. doi:10.1186/1471-2229-8-47
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ficklin, S.P., Feltus, F.A., 2011. Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice. Plant Physiol. 156, 1244-1256. doi:10.1104/pp.111.173047
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L.L., Bateman, A., 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34, D247-D251. doi:10.1093/nar/gkj149
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hansen, B.O., Vaid, N., Musialak-Lange, M., Janowski, M., Mutwil, M., 2014. Elucidating gene function and function evolution through comparison of co-expression networks of plants. Front. Plant Sci. 5, 1-9. doi:10.3389/fpls.2014.00394
Pubmed: Author and Title
CrossRef: Author and Title

Google Scholar: Author Only Title Only Author and Title

He, X., Zhang, J., 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169, 1157-1164. doi:10.1534/genetics.104.037051
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Heiling, S., Schuman, M.C., Schoettner, M., Mukerjee, P., Berger, B., Schneider, B., Jassbi, A.R., Baldwin, I.T., 2010. Jasmonate and ppHsystemin regulate key Malonylation steps in the biosynthesis of 17-Hydroxygeranyllinalool Diterpene Glycosides, an abundant and effective direct defense against herbivores in Nicotiana attenuata. Plant Cell 22, 273-292. doi:10.1105/tpc.109.071449
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Heyndrickx, K.S., Vandepoele, K., 2012. Systematic Identification of Functional Plant Modules through the Integration of Complementary Data Sources. PLANT Physiol. 159, 884-901. doi:10.1104/pp.112.196725
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Hirai, M.Y., Sugiyama, K., Sawada, Y., Tohge, T., Obayashi, T., Suzuki, A., Araki, R., Sakurai, N., Suzuki, H., Aoki, K., Goda, H., Nishizawa, O.I., Shibata, D., Saito, K., 2007. Omics-based identification of Arabidopsis Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc. Natl. Acad. Sci. U. S. A. 104, 6478-6483. doi:10.1073/pnas.0611629104
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Itkin, M., Heinig, U., Tzfadia, O., Bhide, a J., Shinde, B., Cardenas, P.D., Bocobza, S.E., Unger, T., Malitsky, S., Finkers, R., Tikunov, Y., Bovy, A., Chikate, Y., Singh, P., Rogachev, I., Beekwilder, J., Giri, A.P., Aharoni, A., 2013. Biosynthesis of antinutritional alkaloids in solanaceous crops is mediated by clustered genes. Science (80-.). 341, 175-9. doi:10.1126/science.1240230
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Jassbi, A.R., Gase, K., Hettenhausen, C., Schmidt, A., Baldwin, I.T., 2008. Silencing geranylgeranyl diphosphate synthase in Nicotiana attenuata dramatically impairs resistance to tobacco hornworm. Plant Physiol. 146, 974-986. doi:10.1104/pp.107.108811
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2015. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. gkv1070-. doi:10.1093/nar/gkv1070
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Klie, S., Nikoloski, Z., 2012. The choice between MapMan and Gene ontology for automated gene function prediction in plant science. Front. Genet. 3, 1-14. doi:10.3389/fgene.2012.00115
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Kummerfeld, S.K., Teichmann, S. a., 2005. Relative rates of gene fusion and fission in multi-domain proteins. Trends Genet. 21, 25-30. doi:10.1016/j.tig.2004.11.007
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. doi:10.1186/1471-2105-9-559
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J., Shim, J.E., Shim, H., Kim, H., Kim, C., Lee, I., 2015. AraNet v2: an improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. Nucleic Acids Res. 43, D996-1002. doi:10.1093/nar/gku1053
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lee, T.H., Tang, H., Wang, X., Paterson, A.H., 2013. PGDD: A database of gene and genome duplication in plants. Nucleic Acids Res. 41, 1152-1158. doi:10.1093/nar/gks1104
Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Lepiniec, L., Debeaujon, I., Routaboul, J.-M., Baudry, A., Pourcel, L., Nesi, N., Caboche, M., 2006. Genetics and biochemistry of seed flavonoids. Annu. Rev. Plant Biol. 57, 405-430. doi:10.1146/annurev.arplant.57.032905.105252

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Li, S., Ge, F.R., Xu, M., Zhao, X.Y., Huang, G.Q., Zhou, L.Z., Wang, J.G., Kombrink, A., McCormick, S., Zhang, X.S., Zhang, Y., 2013. Arabidopsis COBRA-LIKE 10, a GPI-anchored protein, mediates directional growth of pollen tubes. Plant J. 74, 486-497. doi:10.1111/tpj.12139

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., Van De Peer, Y., 2005. Modeling gene and genome duplications in eukaryotes. Proc. Natl. Acad. Sci. 102, 5454-5459. doi:10.1073/pnas.0501102102

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Mao, L., Van Hemert, J.L., Dash, S., Dickerson, J.A., 2009. Arabidopsis gene co-expression network and its functional modules. BMC Bioinformatics 10, 346. doi:10.1186/1471-2105-10-346

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Matsuno, M., Compagnon, V., Schoch, G. a, Schmitt, M., Debayle, D., Bassard, J.-E., Pollet, B., Hehn, A., Heintz, D., Ullmann, P., Lapierre, C., Bernier, F., Ehlting, J., Werck-Reichhart, D., 2009. Evolution of a novel phenolic pathway for pollen development. Science 325, 1688-1692. doi:10.1126/science.1174095

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

McFarlane, H.E., Döring, A., Persson, S., 2014. The cell biology of cellulose synthesis. Annu. Rev. Plant Biol. 65, 69-94. doi:10.1146/annurev-arplant-050213-040240

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Movahedi, S., Van de Peer, Y., Vandepoele, K., 2011. Comparative network analysis reveals that tissue specificity and gene function are important factors influencing the mode of expression evolution in Arabidopsis and rice. Plant Physiol. 156, 1316-1330. doi:10.1104/pp.111.177865

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Mutwil, M., Klie, S., Tohge, T., Giorgi, F.M., Wilkins, O., Campbell, M.M., Fernie, A.R., Usadel, B., Nikoloski, Z., Persson, S., 2011. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell 23, 895-910. doi:10.1105/tpc.111.083667

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöh, O., Persson, S., 2010. Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. Plant Physiol. 152, 29-43. doi:10.1104/pp.109.145318

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Niggeweg, R., Michael, A.J., Martin, C., 2004. Engineering plants with increased levels of the antioxidant chlorogenic acid. Nat. Biotechnol. 22, 746-754. doi:10.1038/nbt966

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Obayashi, T., Nishida, K., Kasahara, K., Kinoshita, K., 2011. ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. Plant Cell Physiol. 52, 213-219. doi:10.1093/pcp/pcq203

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Onkokesung, N., Gaquerel, E., Kotkar, H., Kaur, H., Baldwin, I.T., Galis, I., 2012. MYB8 Controls Inducible Phenolamide Levels by Activating Three Novel Hydroxycinnamoyl-Coenzyme A:Polyamine Transferases in Nicotiana attenuata. Plant Physiol. 158, 389-407. doi:10.1104/pp.111.187229

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Park, C.Y., Wong, A.K., Greene, C.S., Rowland, J., Guan, Y., Bongo, L.A., Burdine, R.D., Troyanskaya, O.G., 2013. Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. PLoS Comput. Biol. 9, e1002957. doi:10.1371/journal.pcbi.1002957

Pubmed: Author and Title

CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Persson, S., Paredez, A., Carroll, A., Palsdottir, H., Doblin, M., Poindexter, P., Khitrov, N., Auer, M., Somerville, C.R., 2007. Genetic evidence for three unique components in primary cell-wall cellulose synthase complexes in Arabidopsis. Proc. Natl. Acad. Sci. U. S. A. 104, 15566-15571. doi:10.1073/pnas.0706592104

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Persson, S., Wei, H., Milne, J., Page, G.P., Somerville, C.R., 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc. Natl. Acad. Sci. U. S. A. 102, 8633-8. doi:10.1073/pnas.0503392102

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Proost, S., Van Bel, M., Sterck, L., Billiau, K., Van Parys, T., Van de Peer, Y., Vandepoele, K., 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell 21, 3718-3731. doi:10.1105/tpc.109.071506

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, a, Holm, L., Sonnhammer, E., Eddy, S., Bateman, a, Finn, R., 2012. The Pfam protein families databases. Nucleic Acids Res 40 D290-D301. 30, 1-12. doi:10.1093/nar/gkp985

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Rohrmann, J., Tohge, T., Alba, R., Osorio, S., Caldana, C., McQuinn, R., Arvidsson, S., Van Der Merwe, M.J., Riaño-Pachón, D.M., Mueller-Roeber, B., Fei, Z., Nesi, A.N., Giovannoni, J.J., Fernie, A.R., 2011. Combined transcription factor profiling, microarray analysis and metabolite profiling reveals the transcriptional control of metabolic shifts occurring during tomato fruit development. Plant J. 68, 999-1013. doi:10.1111/j.1365-313X.2011.04750.x

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Ruprecht, C., Mutwil, M., Saxe, F., Eder, M., Nikoloski, Z., Persson, S., 2011. Large-Scale Co-Expression Approach to Dissect Secondary Cell Wall Formation Across Plant Species. Front. Plant Sci. 2, 1-13. doi:10.3389/fpls.2011.00023

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Shi, Z., Derow, C.K., Zhang, B., 2010. Co-expression module analysis reveals biological processes, genomic gain, and regulatory mechanisms associated with breast cancer progression. BMC Syst. Biol. 4, 74. doi:10.1186/1752-0509-4-74

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Stuart, J.M., Segal, E., Koller, D., Kim, S.K., 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249-55. doi:10.1126/science.1087447

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tohge, T., Fernie, A.R., 2010. Combining genetic diversity, informatics and metabolomics to facilitate annotation of plant gene function. Nat. Protoc. 5, 1210-1227. doi:10.1038/nprot.2010.82

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tohge, T., Yonekura-Sakakibara, K., Niida, R., Watanabe-Takahashi, A., Saito, K., 2007. Phytochemical genomics in Arabidopsis thaliana: A case study for functional identification of flavonoid biosynthesis genes. Pure Appl. Chem. 79, 811-823. doi:10.1351/pac200779040811

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Tzfadia, O., Amar, D., Bradbury, L.M.T., Wurtzel, E.T., Shamir, R., 2012. The MORPH algorithm: ranking candidate genes for membership in Arabidopsis and tomato pathways. Plant Cell 24, 4389-406. doi:10.1105/tpc.112.104513

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F.M., Bassel, G.W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., Provart, N.J., 2009. Co-expression tools for plant biology: Opportunities for hypothesis generation and caveats. Plant, Cell Environ. 32, 1633-1651. doi:10.1111/j.1365-3040.2009.02040.x

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Wapinski, I., Pfeffer, A., Friedman, N., Regev, A., 2007. Natural history and evolutionary principles of gene duplication in fungi. Nature 449, 54-61. doi:10.1038/nature06107

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yonekura-Sakakibara, K., Tohge, T., Matsuda, F., Nakabayashi, R., Takayama, H., Niida, R., Watanabe-Takahashi, A., Inoue, E., Saito, K., 2008. Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in Arabidopsis. Plant Cell 20, 2160-76. doi:10.1105/tpc.108.058040

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yonekura-Sakakibara, K., Tohge, T., Niida, R., Saito, K., 2007. Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in Arabidopsis by transcriptome coexpression analysis and reverse genetics. J. Biol. Chem. 282, 14932-14941. doi:10.1074/jbc.M611498200

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title

Yu, H., Luscombe, N.M., Qian, J., Gerstein, M., 2003. Genomic analysis of gene expression relationships in transcriptional regulatory networks. Trends Genet. doi:10.1016/S0168-9525(03)00175-6

Pubmed: Author and Title
CrossRef: Author and Title
Google Scholar: Author Only Title Only Author and Title